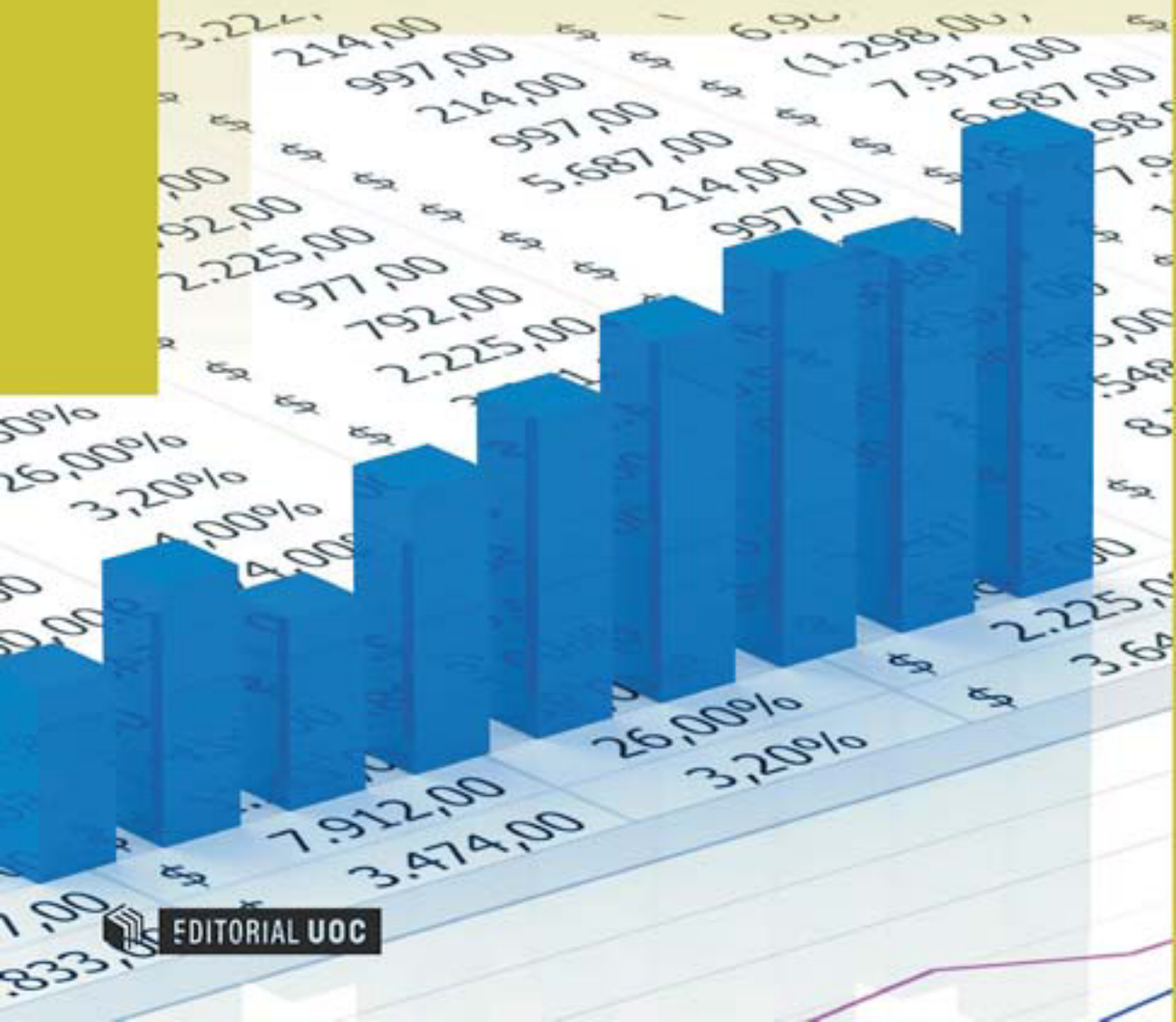
Sociología

Análisis de datos de encuesta

Desarrollo de una investigación completa utilizando SPSS

Vidal Díaz de Rada



EDITORIAL UOC

Análisis de datos de encuesta

Desarrollo de una investigación completa utilizando SPSS

Análisis de datos de encuesta

Desarrollo de una investigación completa utilizando SPSS

Vidal Díaz de Rada

EDITORIAL UOC

Diseño de la coleccion: Editorial UOC

Primera edición en lengua castellana: Junio 2009

© Vidal Díaz de Rada, del texto
© Imagen de la cubierta: Istockphoto
© Editorial UOC, de esta edición
Rambla del Poblenou 156, 08018 Barcelona
www.editorialuoc.com

Los cuadros de diálogo y los listados son copyright de SPSS Inc

Realización editorial: Carrera edició, S.L.
Impresión:

ISBN: 978-84-9788-832-5
Depósito legal

Ninguna parte de esta publicación, incluido el diseño general y la cubierta, puede ser copiada, reproducida, almacenada o transmitida de ninguna forma, ni por ningún medio, sea éste eléctrico, químico, mecánico, óptico, grabación fotocopia, o cualquier otro, sin la previa autorización escrita de los titulares del copyright.

Vidal Díaz de Rada
Licenciado y Doctor en Sociología y profesor de Métodos y Técnicas de Investigación en la Universidad Pública de Navarra. Autor de diversos trabajos sobre metodología, sociología del consumo y comportamiento del consumidor; destacando entre sus últimas publicaciones "Manual del trabajo de Campo en la Encuesta" (2005), "Algunos problemas de la encuesta telefónica para la proyección electoral" (2007) y *Estudio de las incidencias en la investigación mediante encuesta: el caso de los barómetros del CIS* (2008), con A. Núñez Villuendas.

A mi padres,
como hace diez años

Índice

Introducción 13

PARTE I. PROCESO DE INVESTIGACIÓN MEDIANTE ENCUESTAS 17

Capítulo I. Etapas en una investigación con encuestas 19

1. Objetivos didácticos 19
2. Delimitación de objetivos y formulación del problema de investigación 19
3. Elaboración de los objetivos específicos 20
4. Elaboración del cuestionario y proceso de administración 21
5. Prueba del cuestionario 23
6. Construcción de la muestra 24
7. Selección de los entrevistadores 25
8. Formación y entrenamiento del personal seleccionado 25
9. Realización del trabajo de campo y supervisión de las entrevistas 26
10. Codificación de preguntas y depuración de la información 27
11. Tabulación y análisis de datos 27
12. Redacción del informe 28

Capítulo II. Preliminares del análisis de datos 31

1. Objetivos didácticos 31
2. Preparación de los datos para el análisis 31
2.1. Introducción y grabación de la información 35
2.2. Revisión y depuración de la información recogida 36
2.3. Verificación final de la información 39
3. Análisis de una variable 40
4. Relaciones entre dos variables: análisis bivariable 40
5. Análisis multivariable 41
6. Anexo 1: Cuestionario utilizado como ejemplo 42

PARTE II. RECOGIDA Y DEPURACIÓN DE LA INFORMACIÓN 49

Capítulo III. Elaboración de un archivo de datos 51

1. Objetivos didácticos 51
2. Aspectos previos a la introducción de datos: el *formato* de los datos 51
3. Consideraciones previas a la creación de un archivo de datos: menú *vista de variable*s 54
4. Tipos de variables considerando la escala de medida 64
5. Creación de un archivo de datos: definición de variables 68
6. Preparación de los datos: codificación de respuestas 74
7. Introducción de datos (grabación) 82
8. Unión de archivos que contienen una estructura de información similar 84
9. Anexo 1: Libro de códigos (inicial) del cuestionario "Encuestas estudiantes" 88

Capítulo IV. Introducción al SPSS 103

1. Objetivos didácticos 103
2. Contextualización histórica 103
3. Comenzando a trabajar con SPSS. Rutinas elementales de funcionamiento 105
4. Ventana del menú principal 108

Capítulo V. *Importación* de archivos creados por otros 119

1. Objetivos didácticos 119
2. Lectura-recuperación de archivos realizados con hojas de cálculo 120
3. Lectura-recuperación de archivos tipo texto 123
4. Lectura-recuperación de archivos de bases de datos 127

Capítulo VI. Depuración de la informacións 133

1. Objetivos didácticos 133
2. Listado de valores de las variables 134
3. Preguntas filtro y preguntas filtradas 138
4. Comprobación de consistencias lógicas entre variables 142
5. Nivel de representatividad de las respuestas obtenidas 146
6. Ponderar archivo 153

PARTE III. ANÁLISIS DE LA INFORMACIÓN 157

Capítulo VII. La obtención de información 159

1. Objetivos didácticos 159
2. Frecuencias de variables nominales y ordinales 160
3. Resultados de SPSS: visor de resultados y editor de gráficos 167
4. Análisis de respuestas múltiples categóricas 173
5. Análisis de respuestas múltiples dicotómicas 177
6. Estadísticos descriptivos 181
7. Anexo 1: Lenguaje de sintaxis de SPSS con los análisis realizados en la unidad didáctica 184

Capítulo VIII. Transformación de datos y creación de nuevas variables 187

1. Objetivos didácticos 187
2. Recodificación automática 188
3. Agrupación visual 191
4. Recodificar en las mismas variables 202
5. Recodificar en distintas variables 208
6. Cálculos y operaciones: procedimiento calcular 211
7. Creación de nuevas variables uniendo valores en las variables de origen 216
8. Selección de casos mediante criterios condicionales 220
9. Segmentar archivo 225
10. Anexo 1: Lenguaje de sintaxis de los análisis realizados en la unidad didáctica 228

Capítulo IX. Tablas de contingencia con dos variables 233

1. Objetivos didácticos 233
2. Elaboración de tablas de contingencia con dos variables 234
3. Utilización de test estadísticos para conocer la relación entre variables nominales 237
3.1. Relación entre variables utilizando el Chi-Cuadrado 240
3.2. Consideraciones a tener en cuenta en la utilización del Chi-Cuadrado 243
3.3. Estadísticos basados en el Chi-Cuadrado 249
4. Análisis del interior de la tabla 252
4.1. Cálculo y diferencia de porcentajes 252
4.2. Interpretación del interior de la tabla utilizando los residuos 260
5. Utilización de test estadísticos para conocer la relación entre variables ordinales 266
6. Anexo 1: Introducción al cálculo de pares 280
7. Anexo 2: Lenguaje de sintaxis de los análisis realizados en el capítulo 282

Capítulo X. Tablas de contingencia con más de dos variables 289

1. Objetivos didácticos 289
2. Tablas de contingencia de respuestas múltiples categóricas 290
3. Tablas de contingencia de respuestas múltiples dicotómicas 296
4. Relaciones múltiples con tablas de más de dos variables: introducción al análisis multivariante 303
5. Teoría sobre relaciones múltiples 312
6. Anexo 1: Lenguaje de sintaxis de los análisis efectuados 317

Glosario 323

Bibliogafía 329

Introducción

Presentamos una obra en la que se exponen algunas de las técnicas de análisis de datos más habituales en la investigación mediante encuesta, contextualizadas dentro de un proceso de investigación. El objetivo del trabajo no es tanto exponer distintas técnicas de análisis de datos, sino que se centra en cómo llevar a cabo una investigación utilizando determinadas técnicas de análisis de datos. Para ello los distintos capítulos siguen la *secuencia* de una investigación completa, desde la formulación de hipótesis hasta la exposición de resultados e interpretación de los mismos. Este proceso de investigación requerirá utilizar técnicas de análisis de datos, oportunidad que aprovecharemos para exponer el uso y la interpretación de cada una de éstas. Es una obra donde la teoría es utilizada como cimiento de las aplicaciones prácticas.

El elevado número de técnicas de análisis de datos utilizadas en la investigación mediante encuesta aconseja realizar una selección con el fin de explicar algunas de ellas en profundidad. Uno de los criterios más importantes para diferenciar entre las técnicas de análisis de datos hace referencia al tipo de métrica de las variables utilizadas; cualitativas y cuantitativas. El hecho que la mayor parte de la investigación con encuesta utilice variables cualitativas, unido a la escasa atención que éstas han recibido en los manuales de técnicas de análisis de datos, nos anima a centrarnos en las técnicas de análisis de datos para variables cualitativas.

La primera página es el lugar idóneo para realizar algunas consideraciones sobre el contenido de esta obra. Señalar, en primer lugar, que es un libro de *análisis de datos de encuesta* ya que nuestro interés es el planteamiento y posterior desarrollo de una investigación completa, investigación que requiere utilizar procedimientos de análisis de datos, como también precisa la elaboración de unos objetivos concretos, construcción de un cuestionario y prueba del mismo, análisis de los datos recogidos, etc. Dentro de este análisis nuestro interés se centrará en razonar porqué se utiliza una determinada técnica y como interpretar los resultados que proporciona. No nos preocupa la comprensión de fórmulas matemáticas complejas, que además trataremos de evitar en la medida de lo posible[1], sino que deseamos profundizar en la utilización de determinadas técnicas de análisis de datos y en la interpretación de los resultados que proporcionan.

El enorme desarrollo de los paquetes estadísticos, unida a la amplia difusión de ordenadores personales, nos lleva a centrar la atención en los *criterios de utilización* de las téc-

1. "Cada fórmula matemática divide por dos el número de lectores de la obra", señala J. Sevilla Moróder (2005: 17).

nicas de análisis de datos y a la *interpretación de los resultados* proporcionados. En este trabajo se ha elegido uno de los programas estadísticos más utilizados por los investigadores sociales de todo el mundo, el SPSS. Las razones que han motivado esta elección son, entre otras, su elevada difusión, la explicación de determinados procesos estadísticos en sus menús de ayuda, y la existencia de una *versión para estudiantes* que posibilita –a un precio muy asequible– que éstos puedan adquirirlo y familiarizarse con él. Deseamos volver a incidir que no es la explicación del programa estadístico la que guiará la exposición, sino que serán las *necesidades de la investigación* las que aconsejen la utilización –y explicación– de determinados procesos de análisis de datos.

La novedad de este texto, a nuestro juicio, está en la carencia en el mercado de una obra que explique a fondo el proceso completo de una investigación con encuestas. Pese a que en los últimos años se ha producido una proliferación de publicaciones sobre la utilización de programas estadísticos, así como otros textos que utilizan ejemplos donde se exponen los primeros pasos en la investigación con encuesta, el texto que aquí presentamos unifica ambos planteamientos al exponer *problemas* concretos del diseño de la investigación mediante encuesta y las consideraciones que deben tenerse en cuenta a la hora de efectuar el análisis de datos. Creemos que hay una importante carencia en el mercado español de un texto que considere conjuntamente estos planteamientos: por un lado los criterios metodológicos de la investigación con encuesta, por otro el análisis de datos reales de una investigación y, por último, la capacidad de realizar conclusiones y análisis *específicos* utilizando las propiedades de los paquetes estadísticos.

En cuanto a su estructura, este libro está dividido en tres partes. El *proceso de investigación*, que así es como se titula la primera, se centra fundamentalmente en explicar las especificidades del proceso de investigación mediante encuesta. El primer capítulo presenta de forma general las etapas de la investigación, mientras que el segundo se centra específicamente en los procesos relacionados con el análisis de datos. Se trata, de hecho, de un capítulo que sintetiza todo el contenido del libro.

La segunda parte se ocupa de la *recogida y depuración de la información*, contenidos que son explicados en cuatro capítulos (del tercero al sexto). El tercer capítulo está dedicado a la definición de variables y la elaboración del archivo de datos, y contempla aspectos como la medición de los hechos sociales, los tipos de variables empleados para efectuar la medición, codificación de respuestas, creación de un archivo donde serán introducidos los datos, introducción de estos (grabación), y la unión con otros archivos. El objetivo del capítulo es elaborar un archivo de datos con las respuestas a un cuestionario.

En ocasiones los investigadores no introducen los datos sino que *toman* la información de encuestas (o investigaciones) elaboradas por otros colegas. Es lo que se conoce como *investigación secundaria*, que está experimentando un notable auge en los últimos años debido a la gran cantidad de información que ya ha sido recogida y se encuentra a dispo-

sición de los investigadores ("esperando" que alguien la analice). El Centro de Investigaciones Sociológicas (CIS), el Instituto Nacional de Estadística (INE), los institutos de estadística de las comunidades autónomas, etc. recogen diariamente información que ponen a servicio de los investigadores. A este aspecto se dedica el capítulo cinco, a como *recuperar* información elaborada por otros.

Se obtenga la información de una u otra forma (esto es, introduciendo sus propios datos o *tomando* datos de otros), será necesario llevar a cabo unos procesos de depuración de la información obtenida, que son explicados en el sexto capítulo. No se ha hablado del capítulo cuatro, donde se realiza una explicación del SPSS, el software informático que será utilizado para llevar a cabo la investigación. Allí se explica someramente sus *rutinas* de funcionamiento.

Tras la *recogida y depuración de la información* (segunda parte), damos paso a la tercera, centrada específicamente en el *análisis de la información* obtenida. El primer capítulo de esta tercera parte (séptimo capítulo del libro) está dedicado a la presentación univariante de la información, prestando atención especial al análisis de las respuestas múltiples (preguntas multirrespuesta), tan frecuentes en la investigación con encuesta. A continuación un capítulo donde se aborda la transformación de datos y creación de nuevas variables, combinando información de una y varias variables. Los dos siguientes capítulos están dedicados al análisis de tablas de contingencia. En el capítulo nueve se consideran las tablas de contingencia de dos variables, y a continuación las tablas de más de dos variables. El análisis de dos variables genera, en ocasiones, que se realicen afirmaciones sobre relaciones entre variables que –en realidad– no existen, al estar generadas por la influencia de terceras variables (influencias espurias). La localización y eliminación de estas influencias permitirá definir con precisión los factores que inciden en cada una de las relaciones detectadas (capítulo X). Es aquí, sin ninguna duda, donde el lector experimentará más satisfacción al constatar la gran cantidad de información que puede *extraer* de sus encuestas.

El libro está acompañado de una serie de *materiales complementarios* donde se presentan los materiales de esa *supuesta investigación* utilizada para exponer los contenidos del libro. En la *carpeta capítulo 2* se presenta el cuestionario utilizado durante la investigación, junto con la solución de un ejercicio propuestos en el capítulo. La carpeta *capítulo 3* incluye un cuestionario codificado, varios cuestionarios respondidos que deberán ser *grabados* por el lector, así como el archivo de datos utilizado en el libro. Dentro del *capítulo 5* se muestran archivos de datos en otros formatos para que el lector practique la recuperación de información elaborada por otros investigadores; archivos que deberán ser *depurados* en el siguiente capítulo. La *carpeta capítulo 7* incluye otros archivos de datos para que el lector practique y lleve a cabo los aspectos planteados en en el archivo *ejercicios prácticos*. Por último la *carpeta capítulo 9* muestra ejemplos diversos de relaciones entre variables. Cada carpeta incluye también diversos ejercicios con los que practicar lo aprendido.

Se trata de recursos complementarios para que el lector termine DOMINANDO todo el análisis de datos de una investigación mediante encuesta. Son materiales que no se incluyen en el libro, pero que pueden descargarse fácilmente del sitio web de la editorial: www.editorialuoc.com

En cuanto al público para el que se ha escrito de este libro, y aunque en un primer momento fue realizado pensado en utilizarlo como manual en un curso de Análisis de datos en Investigación con Encuesta[2], las sucesivas modificaciones realizadas en los últimos años –tratando de hacer más fácil la compresión del mismo– lo hacen muy apropiado para todo tipo de personas interesadas en la investigación con encuesta: economistas, investigadores de mercado, médicos, sociólogos, trabajadores sociales, psicólogos, etc. El libro está escrito en un lenguaje claro y sencillo, y considera que el lector tiene algunos conocimientos previos sobre investigación social; principalmente sobre muestreo y estadística básica. No obstante, es importante precisar que este libro es para gente que se inicia, no para especialistas en la materia.

En el capítulo de los agradecimientos debemos comenzar reconociendo la enorme paciencia de los estudiantes que han cursado *Informática Aplicada a la Investigación Social* en la Universidad Pública de Navarra, que durante ocho años han soportado las sucesivas *versiones preliminares* de este trabajo. Sus consideraciones han sido de gran ayuda para mejorar la transmisión de conocimientos. Gracias a los compañeros del Departamento de Sociología de la Universidad Pública de Navarra y al Servicio Informático de la universidad, en especial a Eduardo Perales. Ana Díaz, de ADG Estudios de Mercado, corrigió las primeras pruebas del texto en un momento muy difícil para ella. Alba y Mario han sufrido –aún sin saberlo– el proceso de elaboración del libro, aunque contar con la ayuda –y la alegría constante– de Estefanía ha facilitado enormemente las cosas. Este libro está dedicado a mis padres, en un año en el que las cosas no han ido del todo bien.

Vidal Díaz de Rada
Pamplona, mayo de 2009

2. De hecho fueron los propios estudiantes –por peticiones relacionadas durante la actividad profesional– los que me animaron a intentar su publicación.

PARTE I

PROCESO DE INVESTIGACIÓN MEDIANTE ENCUESTAS

Capítulo I

Etapas en una investigación con encuestas

Sara Martínez de Morentin Osés

1. Objetivos didácticos del capítulo

Aunque el fin del libro es realizar una breve introducción a la investigación con encuesta centrada fundamentalmente en las aplicaciones informáticas, consideramos que es una excelente oportunidad para presentar el esquema de una investigación mediante encuesta, las etapas necesarias en un proceso de investigación. La experiencia profesional, y la necesidad pedagógica de exponer de forma pormenorizada y con claridad las tareas a realizar dentro de una investigación con encuestas, nos ha llevado a presentar un esquema dividido en once etapas, al que se dedicará por completo este primer capítulo, y cuyo esquema se presenta en la figura 1.1. Deseamos dejar claro que en este libro tan sólo se realizará un breve comentario de estas etapas, a modo de introducción, remitiendo al lector que precise de una explicación más pormenorizada la lectura de la bibliografía específica sobre el tema.

2. Delimitación de objetivos y formulación del problema de investigación

La primera fase en el diseño de una investigación mediante encuesta comienza con una delimitación clara de las cuestiones o materias a investigar, elaborando un listado de los temas sobre los que se quiere obtener información. El planteamiento del problema debe ser claro y unívoco, exponiendo claramente la naturaleza del problema.

Una de las mejores formas de llevar a cabo esta delimitación es la revisión y análisis de las diversas aportaciones ya realizadas a fin de descubrir cuanto se conoce

sobre ese tema. Grande y Abascal (1999: 34-35) señalan varias estrategias para realizar la delimitación de objetivos y formulación del problema, estrategias que pueden utilizarse de forma aislada o conjunta:

1. Consulta a expertos. Conversaciones y discusiones con las personas que toman decisiones o puedan aportar ideas, con el fin de realizar una puesta en común de conocimientos, valoraciones e inquietudes sobre el problema.
2. Búsqueda y análisis de datos disponibles.
3. Análisis de casos de situaciones similares para conocer cómo se actuó.

3. Elaboración de los objetivos específicos

Esta determinación del tema elegido se concretará en la formulación de los objetivos de la investigación, aunque el objetivo no se define desde la propia investigación sino a partir de lo que se pretende con su realización, es decir, considerando el destino o utilidad de la información recogida. La formulación del problema realizada en la primera fase continuará con la formulación de un *objetivo general*[3] y una serie de *objetivos específicos* donde se concreta el objetivo general a nivel de los diversos aspectos, dimensiones y perspectivas que se desean analizar. Centrados en estos objetivos, es preciso distinguir claramente entre el objetivo general y los específicos:

– El *objetivo general* propone lo que se desea obtener con la investigación planteada. Presenta el enunciado claro y preciso de las metas que se persiguen con esta investigación, delimitando el ámbito temático concreto y la especificación de la población-diana del estudio. Para la consecución del objetivo general será necesario apoyarse en los objetivos específicos.
– Los *objetivos específicos* indican lo que se pretende lograr en cada una de las etapas de la investigación, implicando así un mayor nivel de concreción temporal, temática y estratégica. Es conveniente evaluar estos objetivos en cada paso a fin de conocer los distintos niveles de resultados.

Los objetivos específicos requieren una descomposición del fenómeno a investigar en dimensiones que puedan ser fácilmente medibles, proceso que Padilla et al (1998: 124-126) dividen en dos actividades:

3. En el glosario se presenta una definición de los términos en *cursiva*.

– Clarificar las declaraciones sobre los objetivos del estudio: en esta actividad se "diseccionan" los enunciados sobre el objetivo del cuestionario con el fin de delimitar las áreas de contenido, para posteriormente identificar las variables y aspectos concretos implicados en cada área.
  Un ejemplo ayudará a comprender este proceso. En una investigación que analiza la influencia del nivel cultural en el consumo de unos determinados productos será necesario "diseccionar" el concepto nivel cultural dividiéndolo en determinadas *áreas de contenido* o dimensiones. Por ejemplo nivel de estudios, lectura de libros, lectura de revistas, asistencia a actos culturales, y asistencia al cine.
– Identificar los indicadores necesarios para cada área de contenido. Definidas las *áreas de contenido*, las dimensiones del nivel cultural en el ejemplo anterior, llega el momento de analizar los aspectos concretos implicados en cada una. Con este fin se especifican, para cada una de estas áreas, los aspectos que la componen:
  • Lectura de libros: tipo de libros leídos (novela, best-sellers, libro científico, etc), frecuencia de lectura (número de libros leídos al año), últimos libros leídos.
  • Lectura de revistas: tipo de revistas (científicas, culturales, cotilleos, etc.) y frecuencia de lectura.
  • Asistencia a actos culturales: posibilidad de asistencia en el lugar de residencia, tipo de actos culturales, frecuencia.
  • Asistencia al cine: frecuencia de asistencia, criterios para elegir una película, últimas películas vistas, conocimiento de actores y directores, etc.

El resultado de estas actividades es una relación exhaustiva de los contenidos que el cuestionario debe incluir, que antecede al proceso de realización de las preguntas del cuestionario. La delimitación de los contenidos de la investigación es una tarea lenta y difícil, en la que el investigador debe realizar constantes "vueltas atrás" para asegurarse que han sido recogido todos los conceptos del estudio.

4. Elaboración del cuestionario y procedimiento de administración

Seguidamente se procede a la elaboración del cuestionario mediante la operacionalización de las variables formuladas en la etapa anterior. La elaboración de un cuestionario responde generalmente a tres objetivos: estimar magnitudes, describir una población y verificar hipótesis. En este momento se decide también el procedimiento de administración de la encuesta: personal, telefónica, postal, etc.

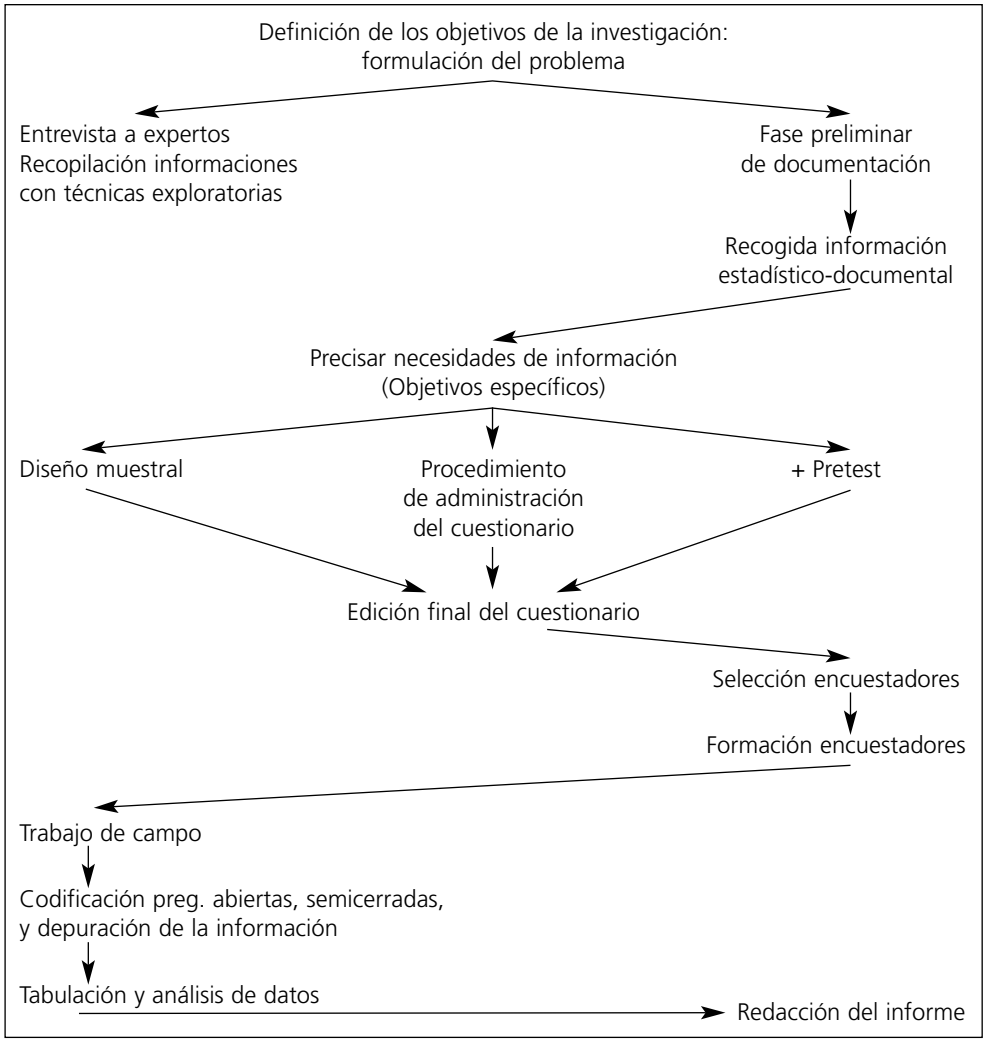


Figura 1.1. Etapas en una investigación mediante encuesta.

La elaboración del cuestionario se ha tratado de forma monográfica en otro texto (Díaz de Rada, 2001). En este libro se utilizarán las respuestas a un cuestionario que ha sido respondido por los estudiantes que cursaron la asignatura Métodos y Técnicas de Investigación Social de la Licenciatura en Sociología de la Universidad Pública de Navarra entre los años 2002 y 2008 (el cuestionario se muestra en el apartado 6). Además, cada estudiante ha pedido a un amigo (que sea estudiante pero que no estudie sociología) que responda este cuestionario. Todas las respuestas, las de los estudiantes de sociología y las de sus amigos, serán utilizadas para hacer prácticas a lo largo de todo el libro con el fin de llegar a conocer los hábitos de los estudiantes de sociología y de sus amigos.

Consideramos que conviene explicar aquí algunas particularidades de este cuestionario, que se ha construido considerando que su objetivo es presentar situaciones que requieran análisis de datos específicos. Es decir, el elemento fundamental en su elaboración es combinar varios tipos de preguntas con el fin de poder utilizar diferentes técnicas de análisis de datos. Por esta razón se combinan preguntas abiertas, cerradas y semiabiertas; se utilizan conjuntamente varios tipos de preguntas de respuesta múltiple (ó multirrespuesta), se proponen determinadas "soluciones" a la codificación de determinadas preguntas, etc. No es frecuente que en un cuestionario se combinen todas estas particularidades, pero el objetivo del libro, aprender las técnicas de análisis de datos, justifica totalmente la decisión tomada[4].

5. Prueba del cuestionario

Definido el cuestionario, es necesario realizar una *prueba* de éste con el objetivo de conocer su adecuación a los objetivos de la investigación; proceso conocido con el nombre "pretest" o "prueba piloto". Para ello el investigador debe realizar varias pruebas del cuestionario no entre sus colegas o familiares, sino con extraños en sus hogares (o en el lugar donde se administre el cuestionario). Una prueba de este tipo suele comprender:

- Averiguar la adecuación de las preguntas realizadas (se entienden, no cometen sesgos, etc.), si el orden del cuestionario es pertinente, y el lenguaje es apropiado para la población objeto de estudio.
- Utilizar diferentes versiones del cuestionario en el que se presentan diversos tipos de preguntas para ver cual funciona mejor.
- Aplicación, mediante entrevista personal, de lo que finalmente será un cuestionario rellenado por el propio entrevistado (autoadministrado), con el fin de detectar problemas de comunicación.
- Formulación de preguntas abiertas que serán utilizadas en la elaboración de categorías de respuesta.
- Diferentes procedimientos de administración del cuestionario, para determinar la adecuación de éstos a los objetivos de la encuesta.

4. Una breve explicación de cada tipo de pregunta, junto con los elementos que les caracterizan, se realiza en Díaz de Rada (2001).

En esta fase, además de realizar una prueba del cuestionario, se "aprovecha la situación" para considerar otros elementos del proceso de investigación. Así Alvira (2004: 19) y Cea D'Ancona (2004: 300-301) señala que –además de probar el cuestionario– el *pretest* suele utilizarse para:

- Analizar la organización del estudio y verificar hasta que punto es adecuada.
- Comprobar la idoneidad del *marco muestral* empleado, verificar la adecuación de la selección muestral efectuada, y conocer el tipo de muestreo seleccionado para, en caso que sea preciso, estimar los parámetros que permitan determinar el tamaño muestral necesario.
- Estimar el porcentaje aproximado de *no respuesta* que se obtendrá en la encuesta, así como la planificación de estrategias para solucionar la no respuesta (revisitas, incentivos, etc.).
- Valorar la adecuación de cómo se realizará la administración del cuestionario (entrevista personal, telefónica, por correo, Internet, etc.). En ocasiones se prueban varios para ver cuál es el que mejor funciona.
- Planificar los contenidos a transmitir en la preparación de los entrevistadores, insistiendo en los elementos más complicados (detectados por el pretest).
- Estimar el tiempo necesario para llevar a cabo el *trabajo de campo*, y así poder realizar una estimación de tiempos y costes con más precisión.

6. Construcción de la muestra

Una vez que el instrumento de medida está terminado y han tenido lugar todos los procesos verificatorios de fiabilidad y validez, llega el momento de la localización de la población de interés que fue definida en la etapa de los objetivos, proceso que comienza con la elaboración de un *marco de muestreo* donde aparecen recogidos todos los elementos de la población. Una vez listada la población objeto de estudio se procede a la selección de una serie de "informantes privilegiados" empleando para ello los procedimientos desarrollados por la teoría muestral.

Algunos expertos sitúan esta etapa antes de la elaboración del cuestionario, mientras que otros proceden con el muestreo después de la elaboración del cuestionario. Es indiferente proceder de una forma u otra, y por este motivo en el esquema de la figura 1.1 están situados al *mismo nivel* el diseño muestral, la elaboración del cuestionario, y la elección del procedimiento de administración del cuestionario. Ahora bien la experiencia investigadora nos ha demostrado que la elabo-

ración previa del cuestionario aclara muchas dudas sobre la población objeto de estudio, y la necesaria representación de determinadas submuestras (cuando proceda).

7. Selección de los entrevistadores

En las encuestas telefónicas y personales el entrevistador es un componente esencial de la recogida de información en la medida que puede influir en la cooperación de los entrevistados y en la calidad de la información recogida. El entrevistador influye en el entrevistado con sus rasgos sociodemográficos, mediante la experiencia obtenida y por las expectativas originadas en la selección de cada entrevistado (Cea D'Ancona, 2004: 309-334): los rasgos del entrevistador afectan a la cooperación en la medida que son evaluados por el posible entrevistado al recibir la visita de alguien que no esperaba. La edad del entrevistador, el género, su lenguaje, etc. serán tenidas en cuenta por el entrevistado en su deseo de conocer realmente el motivo de su visita: vender algo, obtener datos confidenciales de su vida privada, encuestas "encubiertas" que tratan de vender, etc.

La encuesta, definida como la "aplicación de un procedimiento estandarizado para recabar información de una muestra amplia de sujetos", y cuyo objetivo fundamental es la obtención de "mediciones estandarizadas" (Díaz de Rada, 2001: 28), requiere que la administración del cuestionario sea la misma independientemente del entrevistador que la realice. Para ello es necesario llevar a cabo una adecuada selección y formación de los entrevistadores encaminada a reducir las posibles alteraciones generadas por éstos.

8. Formación y entrenamiento del personal seleccionado

Una vez seleccionados los encuestadores llega el momento de proceder con su formación. Aunque en la selección de entrevistadores se ha podido escoger a las personas más capacitadas, es necesario realizar un período de formación en el que deben tratarse todos los aspectos implicados en la actuación del entrevistador. Bajo la premisa que el encuestador no nace, sino que se hace, es preciso diferenciar dos tipos de formación: una formación general donde se transmiten los conocimientos y estrate-

gias básicas para realizar cualquier tipo de encuesta; y una formación específica referida a la investigación que se está desarrollando en ese momento.

Dentro del primer tipo de formación destinada al entrenamiento general de los entrevistadores se tratan aspectos como el proceso de localización del encuestado, la realización de la entrevista, estrategias a utilizar para el manejo de situaciones problemáticas, y cómo realizar una primera revisión de la entrevista. Cuando el *trabajo de campo* se realiza por entrevistadores experimentados, o se confía a una empresa especializada en este tipo de tareas, se puede obviar esta formación general. En cualquier caso, siempre será preciso impartir una formación específica referida a la investigación que se está desarrollando.

9. Realización del trabajo de campo y supervisión de las entrevistas

Terminado el período de formación es el momento de comenzar con la recogida de datos, planificando detalladamente las fechas en las que se realizarán las entrevistas, la labor de los coordinadores de campo, la localización y el horario de la persona a la que acudir cuando aparezcan problemas, etc.

La última tarea de la recogida de información consiste en la *supervisión* y control de las entrevistas. Esta tarea no debe considerarse únicamente como una labor de "control" de los entrevistadores, aunque en las personas que realizan entrevistas por primera vez es conveniente comprobar su grado de honestidad y buena fe, al tiempo que se evalúa su grado de pericia en el desarrollo de este trabajo. De este modo el proceso de supervisión y control ayuda a los entrevistadores a que determinados errores puntuales no se conviertan en hábitos. Así uno de los objetivos que se buscan con esta tarea es detectar aquellos entrevistadores que están trabajando de forma equivocada, aún sin saberlo, y es un proceso que debe intensificarse en los primeros momentos de la recogida de datos para detectar cuanto antes este tipo de errores.

La *supervisión* consiste, básicamente, en la realización de tres tareas: la primera relacionada con la revisión de cuestionarios, la segunda con las incidencias en la recogida de datos y la tercera con la comprobación de las rutas aleatorias. Veamos en detalle cada una de éstas:

1. Revisión de los cuestionarios: comprobación de los datos de identificación, porcentaje de respuestas, codificación de preguntas, omisión de preguntas, calidad de respuestas, seguimiento de preguntas filtro, etc. Esta tarea será muy sencilla si el entrevistador ha realizado correctamente la *revisión de la entrevista*.

2. Conversar con los jefes de equipo y zona, así como con los entrevistadores para conocer las incidencias ocurridas en la administración del cuestionario.
3. Cuando las unidades últimas del muestreo se han elegido mediante rutas aleatorias, el supervisor debe examinar las rutas sobre el terreno.

10. Codificación de preguntas y depuración de la información

Una vez finalizado el proceso de recogida de la información tiene lugar una tarea de revisión de los cuestionarios, donde se realiza una inspección y corrección de las respuestas. La primera revisión se realiza inmediatamente después de realizar la entrevista, durante la realización del *trabajo de campo*, y en ella se repasan todas las preguntas y se comprueba que no se olvidó ninguna, se buscan *contradicciones lógicas* en las respuestas, etc. tratando también de facilitar y asegurar la comprensión de las preguntas para los codificadores.

Una segunda revisión se realiza cuando los cuestionarios han llegado al instituto de investigación, y consiste en la precodificación de la información difícil (ocupaciones, preguntas semi-abiertas, etc.), así como la preparación de resúmenes para simplificar el proceso de codificación de las preguntas abiertas.

La creación de un fichero de datos y la grabación de las respuestas en el mismo son tareas que se realizan a continuación, y que serán analizadas pormenorizadamente en el tercer capítulo. Cuando todos los cuestionarios han sido introducidos en el ordenador es necesario realizar unos proceso de revisión y depuración de los datos con el objetivo de evaluar –y si es posible aumentar– la calidad de la información recogida (lo veremos en el capítulo VI). Se trata de buscar *inconsistencias* entre ciertas preguntas, verificar si hay valores que no tienen lugar en determinadas preguntas, analizar las respuestas de las preguntas filtro, cuantificar la *no respuesta* y decidir que hacer con ella, etc.

11. Tabulación y análisis de datos

Finalizada la fase de preparación de la información comienza la etapa de análisis de datos. En un primer momento el objetivo se centra en obtener un conocimiento detallado de cada una de las variables utilizadas en la investigación, empleando para ello distribuciones de frecuencias, estadísticos univariantes y representaciones gráficas.

Una *distribución de frecuencias* es una tabla donde se muestran las elecciones de las distintas categorías que componen la variable. Cuando la variable se ha medido a nivel de intervalo es aconsejable la utilización de *estadísticos* que además permiten presentar la información en un formato más reducido. El análisis y la presentación de la información mejora notablemente cuando se utilizan *gráficos* para la presentación de los resultados.

El análisis de una variable permite un primer conocimiento de la realidad objeto de estudio, además de preparar los datos para que puedan ser utilizados en las relaciones bivariantes. Aunque el primer conocimiento de la realidad obtenido mediante el análisis univariante es un paso previo e imprescindible antes de proceder con las relaciones entre variables, una mayor riqueza de análisis presenta la utilización de diversas técnicas de análisis bivariante o multivariante. La elección entre cada una de estas técnicas requiere, en primer lugar, considerar la escala en la que se ha medido cada variable (nominal, ordinal, intervalo o razón), y en segundo lugar delimitar claramente la relación que se establece entre las variables: diferencia de porcentajes, diferencia de medias, relación significativa entre variables, etc. Un análisis detallado puede verse en Díaz de Rada (1999).

12. Redacción del informe

La labor de tratamiento y análisis de la información precede a la elaboración de un *informe* donde se presentan los hallazgos de la investigación. Este proceso de presentación y difusión de la información es uno de los aspectos más importantes de la investigación, en la medida que es la única forma de dar a conocer los resultados al demandante de la misma, a los medios de comunicación y a la comunidad científica.

El informe de investigación es, en definitiva, el trabajo de investigación en sí mismo puesto que dentro del informe se explican los objetivos de la investigación, los métodos y técnicas utilizadas para lograr los objetivos, así como las conclusiones del trabajo. De modo que el informe de investigación no sólo muestra los resultados de esta investigación concreta, sino que puede ser utilizado para que otros investigadores repitan esta investigación en otros contextos, o para probar la adecuación de la metodología utilizada para resolver problemas similares. Un buen informe debe desempeñar tres objetivos: reunir lo más importante de una investigación e integrar los resultados en un marco teórico; acumular ciencia al comparar sus resultados con los de otras investigaciones y calibrar su validez; y ayudar a descubrir las lagunas que quedan y los nuevos interrogantes que se plantean, al tiempo que contribuye a aclarar los puntos que han quedado oscuros.

El código deontológico de ESOMAR[5] (2000) señala que cualquier informe o interpretación de una encuesta debe incluir –al menos– la siguiente información: nombre del instituto que ha realizado el sondeo, universo representado, tamaño de la muestra lograda y su extensión geográfica, fecha del trabajo de campo, método de muestreo y tasa de respuesta, procedimiento de campo utilizado, y las preguntas relevantes realizadas (escribiéndolas de un modo literal). Además, el instituto tiene que facilitar al cliente para el que lleva a cabo la encuesta la siguiente información[6]:

1. Antecedentes: para quién se realiza el estudio; finalidad del mismo; y nombre de los subcontratados y consultores que realicen parte del trabajo.
2. Muestra: descripción del universo cubierto (proyectado y real); tamaño, sustituciones y distribución geográfica de la muestra (proyectada y conseguida); qué datos fueron obtenidos sólo de una parte de la muestra; método de muestreo y métodos de ponderación utilizados; informe de tasas de respuestas; e informes de posibles sesgos debidos a la no respuesta.
3. Recogida de datos: descripción del procedimiento de campo; descripción del personal de campo; instrucciones dadas y métodos de control de la calidad del campo; método de reclutamiento/selección de los entrevistados; incentivos ofrecidos para lograr colaboración; y fechas del campo.
4. Presentación de resultados: Resultados reales relevantes obtenidos; bases de los porcentajes (tanto ponderados como no ponderados); indicadores del margen de error estadístico y probabilidades de los resultados principales; medidas de significación estadística de las diferencias entre cifras clave; y el cuestionario y otros elementos y materiales relevantes utilizados.

Esperamos que el lector no se abrume después de haber leído hasta aquí y comprobar la gran cantidad de tareas que se llevan a cabo en cada etapa de la investigación mediante encuestas. El objetivo de este capítulo no es otro que ofrecer una *visión panorámica* del proceso de investigación, presentado sintéticamente todos los conocimientos precisos para llevar a cabo una investigación de este tipo. El objetivo de este libro es más específico, concretamente realizar un primer análisis de los datos (etapas 10 y 11).

5. European Society for Opinion and Marketing Research Association, Asociación Europea de Estudios de Mercado y Opinión.
6. Reproducido de Alvira, 2004: 94-95.

Capítulo II

Preliminares al análisis de datos

1. Objetivos didácticos del capítulo

En el presente capítulo se exponen los aspectos preliminares al análisis de datos –aspectos que se explicaron brevemente en las etapas 10 y 11 de el anterior capítulo– y su objetivo es *situar* al lector en estas tareas del proceso de investigación. Digamos que es un breve resumen de todo el texto, un resumen centrado en los conceptos y procesos a seguir a la hora de realizar una investigación con encuesta, y que deja de lado las instrucciones y procedimientos del programa SPSS; sobre los que profundizaremos en el resto de capítulos. Se trata, en definitiva, de *contextualizar* al lector en el proceso de investigación, destacando las diversas fases del proceso de análisis de datos (últimas etapas de la figura 1.1). Así, mientras que en el primer capítulo se presentó todo el esquema de investigación, en éste se describen las acciones que deben realizarse para el procesamiento de la información.

Es preciso insistir en que el punto de partida es un investigador que ha recogido información de una determinada realidad, y por este motivo la exposición comienza presentando los diversos procedimientos a seguir a la hora de *preparar* los datos para el análisis (apartado dos). Los apartados tres, cuatro y cinco están dedicados al análisis de la información, donde se realiza una somera explicación del análisis de una variable (*univariable*), de dos variables al mismo tiempo (*bivariable*), y de más variables simultáneamente (*multivariable*).

2. Preparación de los datos para el análisis

La etapa de tratamiento de la información comienza en el momento que finaliza el proceso de recogida (o producción) de la información, aunque en la práctica existen unos *solapamientos temporales* puesto que una de las primeras fases del tratamiento de la información es la revisión de los cuestionarios (o cualquier otro formato elegido para la recogida de la información), y esto se lleva a cabo durante la realiza-

ción del trabajo de campo, inmediatamente después de responder al cuestionario (recordar capítulo anterior). Debe quedar claro que en términos de programación de la investigación ambas son fases independientes, aunque la revisión de cuestionarios se realice –por motivos prácticos– justo después de la entrevista[7]. La revisión realizada después de la entrevista permite corregir fácilmente los errores producidos en la cumplimentación del cuestionario, bien preguntando al entrevistador unas horas después de realizar la entrevista, o incluso volviendo a solicitar la cooperación del entrevistado para algunos aspectos puntuales de la entrevista. Esta primera corrección de errores es muy efectiva unas horas después de la entrevista, mientras que no serviría para mucho si se hiciera unos días después.

Este proceso de primera revisión de los cuestionario es conocido con el nombre de *edición*, y puede ser definido como "el proceso en el que las respuestas son inspeccionadas, corregidas y en ocasiones precodificadas de acuerdo a un conjunto de reglas fijas" (Díaz de Rada, 1999: 72). Tras la inspección y corrección de las respuestas llega el momento de la elaboración del archivo de datos.

Sin embargo, antes de proceder con la elaboración del el archivo de datos hay que *trasladar* las respuestas del cuestionario a información analizable, almacenando esta información mediante una representación numérica o por medio de otros símbolos, proceso conocido con el nombre de *codificación*. El proceso de codificación se compone de dos tareas: en un primer momento se introducen códigos en el cuestionario (o en el formato elegido para la recogida de la información) con el fin de identificar cada pregunta con un nombre determinado, para proceder seguidamente con la elaboración de un *libro de códigos*[8] donde se recogen todas las opciones o categorías de respuestas. La mayoría de definiciones de codificación se refieren a este último proceso. Este es el caso, por ejemplo, de Alvira (2004: 54) cuando define *codificar* como clasificar de un modo sistemático en categorías mutuamente excluyentes y exhaustivas las respuestas a las preguntas para poder después transferir la información a un soporte informático y analizarlas".

En el cuadro 2.1 se presenta un ejemplo de codificación de una investigación realizada para conocer las características de las personas que pertenecen a una asocia-

---

7. Reproducimos de forma literal una cita de Francisco Alvira (2004: 43) donde da cuenta de este proceso:

   "En la encuesta personal se desarrollan en paralelo los procesos de supervisión de la tarea de los encuestadores, depuración y codificación de los cuestionarios"... lo que implica que es preciso una organización compleja de técnicos: entrevistadores, supervisores y codificadores/depuradores". "Cada cuestionario, una vez cumplimentado por el entrevistador, pasa por la consiguiente *revisión* y *supervisión* (esta última sólo en un porcentaje que oscila entre el 15% y el 25%) y con posterioridad es *depurado* y en su caso, *codificado*".

8. Tal y como se señaló en la nota a pie número 3 del primer capítulo, en el glosario se presenta una definición de los términos en *cursiva*.

ción que presta trabajo voluntario (en la figura 2.1 se muestran las fichas de cada voluntario). En un primer momento se identifica cada una de las informaciones recogidas con un nombre, para especificar a continuación sus categorías de respuesta. La mayor

FICHA DE AFILIACIÓN A LA ASOCIACIÓN

Nombre: *Carmen* Sexo: *Mujer*

Horas de trabajo a la semana: *15 horas*
Tipo de jornada: *Mañana*
Horas disponibles (semanales): *30*
¿Colabora con otras asociaciones? *No*
Tipo de asociaciones con las que colabora:

FICHA DE AFILIACIÓN A LA ASOCIACIÓN

Nombre: *Juan* Sexo: *Hombre*

Horas de trabajo a la semana: *35 horas*
Tipo de jornada: *Mañana y tarde*
Horas disponibles (semanales): *5*
¿Colabora con otras asociaciones? *Si*
Tipo de asociaciones con que colabora:
Recreativas

FICHA DE AFILIACIÓN A LA ASOCIACIÓN

Nombre: *Iker* Sexo: *Varón*

Horas de trabajo a la semana: *40 horas*
Tipo de jornada: *Partida*
Horas disponibles (semanales): *7*
¿Colabora con otras asociaciones? *Si*
Tipo de asociaciones con las que colabora:
Deportivas

FICHA DE AFILIACIÓN A LA ASOCIACIÓN

Nombre: *Nacho* Sexo: *Hombre*

Horas de trabajo a la semana: *35 horas*
Tipo de jornada: *Sólo mañanas*
Horas disponibles (semanales): *10*
¿Colabora con otras asociaciones? *Si*
Tipo de asociaciones con que colabora:
Culturales y deportivas

FICHA DE AFILIACIÓN A LA ASOCIACIÓN

Nombre: *Oiana* Sexo: *Mujer*

Horas de trabajo a la semana: *30 horas*
Tipo de jornada: *Partida*
Horas disponibles (semanales): *15*
¿Colabora con otras asociaciones? *Si*
Tipo de asociaciones con las que colabora:
Deportivas, recreativas y gastronómicas

FICHA DE AFILIACIÓN A LA ASOCIACIÓN

Nombre: *Olatz* Sexo: *Mujer*

Horas de trabajo a la semana: *40 horas*
Tipo de jornada: *Continua*
Horas disponibles (semanales): *12*
¿Colabora con otras asociaciones? *Si*
Tipo de asociaciones con que colabora:
Recreativas

FICHA DE AFILIACIÓN A LA ASOCIACIÓN

Nombre: *Pedro* Sexo: *Varón*

Horas de trabajo a la semana: *35 horas*
Tipo de jornada: *Mañana y tarde*
Horas disponibles (semanales): *20*
¿Colabora con otras asociaciones? *Si*
Tipo de asociaciones con las que colabora:
Deportivas y gastronómicas

FICHA DE AFILIACIÓN A LA ASOCIACIÓN

Nombre: *Víctor* Sexo: *Hombre*

Horas de trabajo a la semana: *50 horas*
Tipo de jornada: *Continua*
Horas disponibles (semanales): *3*
¿Colabora con otras asociaciones? *No*
Tipo de asociaciones con que colabora:

Figura 2.1. Información de los afiliados a la asociación ASOCUCAR. (Fichas de afiliación).

FASE 1: IDENTIFICACIÓN DE CADA PREGUNTA

Identificación cada pregunta	Significado
Nombre:	Nombre del voluntario.
Sexo:	Sexo.
H_trab:	Promedio de horas trabajadas cada semana
Jornada:	Tipo de jornada
H_dispo:	Horas disponibles para emplear en tareas de cooperación (por semana)
N_asoc:	Pertenencia a asociaciones: número.
T_asoc:	Pertenencia a asociaciones: tipo.

FASE 2: OPCIONES DE RESPUESTA DE CADA PREGUNTA

Nombre del voluntario:
- Nombre.

Sexo:
- Varón (V).
- Mujer (M).

Promedio de horas trabajadas cada semana:
- Número horas.

Tipo de jornada:
- Partida (1).
- Continua (2).

Horas disponibles para emplear en tareas cooperación:
- Número horas.

Pertenencia a asociaciones: número.
- Número asociaciones a las que pertenece.

Pertenencia a asociaciones: tipo.
- Culturales (1).
- Deportivas (2).
- Recreativas (3).
- Gastronómicas (4).

LIBRO DE CÓDIGOS DONDE SE SINTETIZA TODA LA INFORMACIÓN ANTERIOR

Número Pregunta	Nombre	Códigos	Etiquetas
1	Nombre		Nombre del voluntario
2	Sexo		Sexo
		V	Varón
		M	Mujer
3	H_trab		Promedio de horas trabajadas cada semana
		—	Número de horas
4	Jornada		Tipo de jornada
		1	Jornada partida (mañana y tarde).
		2	Jornada continua (mañana ó tarde)
5	H_dispo		Horas disponibles para emplear en tareas de cooperación (por semana)
		—	Número de horas
6	N_asoc		Número de asociaciones que pertenece
		—	Número de asociaciones
7	T_asoc		Tipo de asociaciones
		1	Culturales.
		2	Deportivas.
		3	Recreativas.
		4	Gastronómicas.

Cuadro 2.1. Ejemplo de codificación.

Nombre	Sexo	H_trab	Jornada	H_dispo	N_asoc	T_asoc
Carmen	M	15	2	30	0	-
Juana	V[10]	35	1	5	1	3
Iker	V	40	1	7	1	2
Nacho	V	35	2	10	2	1,2
Oiana	M	30	1	15	3	2,3,4
Olatz	M	40	2	12	1	3
Pedro	V	35	1	20	2	2,4
Victor	V	50	2	3	0	-

Cuadro 2.2. Ejemplo de matriz de datos.

parte de las veces esta información aparece sintetizada en un *libro de códigos* que acompañara al investigador durante toda la investigación[9].

La creación del archivo de datos consiste en la elaboración de una *matriz* de datos de *n* filas por *m* columnas donde queda recogida la información de los entrevistados en cada una de las preguntas del cuestionario. Un fichero de datos puede ser muy simple, unas pocas filas y columnas de números, o contar con una elevada cantidad de datos, definición de variables, *filtros* de preguntas y especificaciones de *depuración* de la información. En la matriz de datos la opinión de cada unidad analizada (cada entrevistado) queda situada en una fila, mientras que en las columnas aparece la información de cada una de las variables medidas; de modo que la unión de una fila y columna es la respuesta de una determinada persona a una pregunta. En el cuadro 2.2 se muestra un ejemplo de una matriz de datos, concretamente la información del ejemplo sobre las personas que pertenecen a una asociación de voluntariado, y cuyos valores aparecieron en el cuadro 2.1.

Existe también la posibilidad de introducir los datos en un fichero estándar tipo texto, por ejemplo utilizando el *formato ASCII*, y después realizar el proceso de definición de las variables mediante el programa estadístico que se utilizará para la lectura y análisis de los datos, aspectos que desarrollaremos en el capítulo V.

2.1. Introducción y grabación de la información

Una vez que las variables han sido definidas se procede a introducir las respuestas de los cuestionarios al archivo de datos. La mayor parte de las veces los investiga-

9. Profundizaremos más en este aspecto en el capítulo III, donde se introducen los datos del cuestionario presentado al final del presente capítulo.
10. Se trata de un error puesto que Juana no puede codificarse con "v" de varón (ver cuadro 2.1). Esto justifica tener que revisar la matriz de datos.

dores utilizan los subprogramas de grabación de la información que acompañan a los paquetes estadísticos, o bien emplean los programas de bases de datos que se encuentran más accesibles en el mercado. Debe tenerse en cuenta que el proceso de grabación de la información lleva consigo un chequeo de los valores introducidos y –cuando sea pertinente– un cambio de los mismos.

Todos los investigadores están de acuerdo que la tarea de introducir datos es aburrida y tediosa, factores que originan la aparición de un gran número de errores en el proceso de introducción de datos. El desarrollo y aplicación de nuevas tecnologías ha generado que muchas organizaciones dediquen actualmente una gran cantidad de recursos a la investigación sobre formas *automáticas* de introducción de datos, que además de ser más rápidas reducen la posibilidad de errores. Los avances producidos en los últimos años parecen augurar que nombres como *OCR* (reconocimiento óptico de caracteres), *OMR* (reconocimiento mecánico de caracteres) y otros sistemas de grabación óptica serán familiares para nosotros dentro de unos años (Mejías, 2005: 5).

2.2. Revisión y depuración de la información recogida

Cuando todos los cuestionarios han sido introducidos en el ordenador es necesario realizar un proceso de revisión y *depuraci*ón de los datos con el objetivo de evaluar –y si es posible aumentar– la calidad de la información recogida. Se trata de buscar inconsistencias entre ciertas preguntas, verificar si hay valores que no tienen lugar en determinadas preguntas, analizar las respuestas de las *preguntas filtro*, cuantificar la *no respuesta parcial* y decidir que hacer con ella, etc. Pese a la importancia de esta labor, la revisión y *depuraci*ón de la información recogida cada día recibe menos atención debido a la premura con la que se realiza la recogida y análisis de resultados, la rutina de estas tareas y el abuso de los ordenadores (Villán y Bravo, 1990: 15). Estos factores han generado un descuido en los trabajos de depuración, olvidando el enorme impacto que tienen en la calidad de los datos recogidos y, en última instancia, en la calidad en la investigación.

Adoptaremos la definición que utiliza Félix Aparicio cuando define *depuración de datos* como "un conjunto de técnicas que permiten, a partir de la información recogida en la encuesta, y a veces a partir de otra información adicional, corregir una parte de los errores de la encuesta" (Aparicio, 1991: 92). El momento temporal en el que se lleva a cabo el proceso de verificación y depuración de la información estará condicionado por el procedimiento de recogida de datos, ya que algunos permiten verificar la información cuando el entrevistado responde el cuestionario: en entrevistas personales asistidas por ordenador (*Computer Assisted Personal Interview-CAPI*), entrevistas telefónicas asistidas por ordenador (*Computer Assisted Telephone Interview-CATI*) o encues-

tas autorellenadas asistidas por ordenador (*Computer Assisted Self Interviewing-CASI* y *Computer Assisted Web Interviewing-CAWI*) las respuestas de cada entrevistado son grabadas en el mismo momento en que responden. La creación de filtros y otros instrumentos permitirá realizar la labor de edición en el momento de la entrevista, posibilitando inmediatamente la localización de *contradicciones lógicas*, permitiendo así la repetición de ciertas preguntas al entrevistado. Supongamos una entrevista en la que una mujer señala que tiene 13 años y afirma tener 15 hijos. Cuando esto ocurre en los sistemas CATI, CAPI, CASI y CAWI el software avisa al entrevistador de esta inconsistencia, permitiendo rápidamente la repetición de ambas preguntas.

Cuando la investigación se realiza mediante entrevistas personales con cuestionario de papel el proceso de revisión y depuración tiene lugar después de la grabación de la información. En numerosas ocasiones esta revisión se lleva a cabo en el momento mismo en que son introducidos los datos, realizando la introducción de datos por duplicado y empleando personas distintas (Granero et al, 2001: 1-13). Posteriormente se comparan ambos ficheros a fin de detectar las diferencias existentes; proceso conocido como *depuración por contraste*. Las inconsistencias se solucionarán consultado los cuestionarios originales.

En el momento que se dispone de la información en formato magnético es posible localizar de forma rápida y eficaz los fallos cometidos durante la recogida y grabación de datos. En los siguientes párrafos se exponen algunos de los sistemas de validación más utilizados en el proceso de revisión y depuración de la información recogida[11]:

1. El primero se fundamenta en la petición de un listado de los valores de todas las variables del cuestionario, realizando tabulaciones para cada variable. La *distribución de frecuencias* resultante de comparan con las tarjetas del *libro de códigos* con el objetivo de comprobar si alguna de ellas tiene valores ajenos al recorrido de la variable, o valores que no aparecen en el libro de códigos. Por ejemplo un valor "7" en la pregunta 24 (sexo) del cuestionario mostrado en el apartado 2.6. El sexo únicamente presenta dos posibilidades; hombre (codificado con el valor 1) y mujer (codificado con el valor 2), de modo que un "7" será un error.

   Cuando esto sucede se procede a buscar en el fichero de datos el número de caso donde aparece este valor, para localizar a continuación el cuestionario original. Cuando es un error en el proceso de grabación de la información basta con cambiarlo por el valor verdadero.

   Una situación más problemática surge cuando el cuestionario está mal respondido, en cuya situación nada puede hacerse para mejorar la calidad de este

11. Veremos más adelante, en el capítulo VI, cómo proceder con el programa estadístico elegido cuando el investigador se encuentre con esa situación.

registro. Cuando se realiza una correcta labor de edición y los cuestionarios son revisados por los entrevistadores y los coordinadores de campo es muy difícil que se produzca esta situación.

Por otro lado, los modernos programas de grabación de datos permiten delimitar los valores máximos y mínimos de las variables, de modo que es prácticamente imposible introducir datos fuera de rango.

2. Una vez que se ha verificado que todos los valores se ajustan al recorrido de las variables, en la segunda comprobación se comparan el número de respuestas de las *preguntas filtro* con las *preguntas filtradas*[12].

   En el cuestionario del apartado 2.6 esto implicaría que las personas que no tienen vídeo o DVD (pregunta 16, opción 2) no deben responder las preguntas 16a, 16b, 16c y 16d referidas al número de películas vistas, día de la semana en que visionó la última película, forma de visionado (sólo o acompañado, etc.). Eso mismo cabe decir de la pregunta 17a, que únicamente es respondida por los que declaran tener ordenador en su hogar, opción 1 de la pregunta 17.
3. El tercer procedimiento consiste en la elaboración de *consistencias lógicas* (o *relaciones lógicas*) que deben ser cumplidas por determinadas variables del cuestionario. En la localización de estas consistencias hay que diferenciar entre "respuestas inconsistentes" y "respuestas improbables": las primeras se refieren a situaciones que es imposible que se cumplan, mientras que las segundas muestran respuestas que son posibles, pero muy improbables. La situación referida anteriormente, una mujer de 13 años que declara tener 15 hijos, es un ejemplo de respuesta inconsistente. Si la persona que afirma tener 15 hijos es una mujer de 27 años se trata de una respuesta improbable, puesto que es posible que suceda; aunque es bastante improbable que esta mujer haya tenido su primer embarazo a los 14 años, y que tenga un hijo cada 10 meses.
4. También debe considerarse el nivel de representatividad de las respuestas obtenidas, analizando la tasa de respuesta de cada pregunta y los niveles de respuesta de cada sujeto.

Esta fase de revisión y depuración de la información termina con la realización de un primer análisis descriptivo de los datos con el objetivo de conocer los *valores atípicos* presentes en la matriz de datos; definidos como observaciones que muestran inconsistencias con el resto de la distribución.

12. Otros autores, por ejemplo Alvira (2004: 28), se refieren a éstas como *preguntas contingentes*.

2.3. Verificación final de la información

En este último paso de la preparación se considera la valoración de la *no respuesta parcial* y *no respuesta total*, la realización de transformaciones de los datos, y el empleo de ponderaciones para corregir problemas de representación de la muestra. El elemento definitorio de esta etapa es que esta verificación tiene lugar tras la primera tabulación de las variables del cuestionario.

Ante la presencia de la no respuesta será conveniente comparar los rasgos sociodemográficos de los que no responden a fin de conocer si la no respuesta es aleatoria, o si más bien se produce en unos estratos determinados. Diversas investigaciones han negado que la no respuesta sea aleatoria (entre otros Díaz de Rada 2000 y Sánchez Carrión 2000), de modo que el principal problema que presenta la no respuesta es la introducción de sesgos a la hora de realizar inferencias.

El proceso de preparación de los datos para el análisis prosigue considerando la conveniencia de emplear coeficientes de ponderación con el fin de *devolver* a cada estrato su *peso* proporcional cuando se han empleado muestreos con afijaciones diferentes a la proporcional; o utilizar la ponderación para que determinados colectivos adopten un *peso* muestral superior al que tienen en la población (se verá con detalle en el capítulo VI).

Para finalizar, es preciso realizar diversas *transformaciones* en algunas variables, transformaciones que van desde la simple unión de varias categorías con contenidos similares, a transformaciones que persiguen la consecución de una distribución aproximada a la normal, simétrica, etc. Otro tipo más complejo de transformaciones es la elaboración de *índices* creados mediante la combinación de determinadas preguntas o variables[13]; aspectos que serán analizados en detalle en el capítulo VIII (apartados 6 y 7). En su libro sobre la *perspectiva general metodológica de la encuesta*, Francisco Alvira (2004: 58) esgrime tres razones que justifican la generación de nuevas variables:

- "No suele existir una correspondencia unívoca entre conceptos y preguntas en la totalidad de las preguntas de una encuesta.
- Reformular y alterar la información básica obtenida resulta a veces imprescindible para poder aplicar determinadas técnicas de análisis estadístico.
- A veces hay que añadir nuevas variables que no se basan en información recogida durante el campo".

13. Por ejemplo clase social, nivel cultural, etc.

3. Análisis de una variable

Terminada la fase de preparación de la información comienza el análisis de los datos que, como tendremos ocasión de profundizar más adelante, viene determinado por la técnica de recogida de información utilizada y por los objetivos de la investigación. El fin en esta primera fase de análisis es obtener un conocimiento detallado de cada una de las variables utilizadas en la investigación, empleando para ello distribuciones de frecuencias, estadísticos univariantes y representaciones gráficas.

El análisis de los datos debe seguir una *línea jerárquica ascendente* que se inicia con la descripción ó exploración de la información, para continuar con el análisis de relación entre variables. Entre las técnicas más utilizadas para el *análisis univariable* destaca la *distribución de frecuencias*, que es una tabla donde se exhiben los distintos valores que componen la variable. Las distribuciones de frecuencias se emplean normalmente para los niveles más bajos de medición, si bien puede ser utilizada para conocer la distribución de valores de cualquier tipo de variable.

Cuando la variable se ha medido a nivel de intervalo es aconsejable la utilización de determinados *estadísticos* que, utilizando distribuciones gráficas, presentan la información en un formato más reducido. Los estadísticos que presenta el SPSS están distribuidos en cuatro grandes grupos: a) Medidas de tendencia central donde se incluye la media, la mediana y la moda; b) medidas de dispersión: desviación típica, varianza y rango; c) medidas de la forma de la distribución: asimetría y curtosis; y d) otras medidas como el valor mínimo, valor máximo y la suma de valores.

El análisis y la presentación de la información mejora notablemente cuando se utilizan *gráficos* para la presentación de los resultados: entre éstos destacamos el uso de diagramas de barras, histogramas, gráficos de tronco y hojas (*stem-and-leaf*), gráfico de caja y bigotes (*box-plot*), polígonos de frecuencias, ojivas, gráficos de sectores, etc.

4. Relaciones entre dos variables: análisis bivariable

El análisis de una variable permite un primer conocimiento de la realidad objeto de estudio, además de *preparar* los datos para que puedan ser utilizados en las relaciones bivariantes. Este primer conocimiento de la realidad obtenido mediante el análisis univariante es un paso previo e imprescindible antes de proceder con las relaciones entre variables.

En la investigación con encuesta las técnicas bivariantes más utilizadas son el análisis de la correlación lineal y el cruce de tablas (o tablas de contingencia) entre dos o más variables. Determinadas situaciones precisan el empleo de otras técnicas como la diferencia significativa de medias, el análisis de varianza, la regresión simple, y los test no paramétricos. La elección entre cada una de estas técnicas requiere, en primer lugar, considerar la métrica en la que se ha medido cada variable (nominal, ordinal, intervalo o razón[14]) y, en segundo lugar, delimitar claramente la relación que se establece entre las variables: diferencia de porcentajes, diferencia de medias, relación significativa entre variables, etc.

En este texto, cuyo objetivo es realizar una introducción al análisis de datos, analizaremos someramente el cruce de tablas (o tablas de contingencia), por los motivos que esgrimimos en la introducción A los interesados en el resto de técnicas les recomendamos la lectura de otros textos (por ejemplo Abascal y Grande 2005, Díaz de Rada 1999, Sánchez Carrión 1999).

5. Análisis multivariable

La complejidad de los fenómenos sociales obliga a los investigadores a recoger una gran cantidad de medidas con el fin de captar de forma adecuada la naturaleza de los fenómenos analizados. Esto genera que, en numerosas ocasiones, los *análisis univariables y bivariables* sean insuficientes para resolver adecuadamente los objetivos de la investigación, por su imposibilidad para proporcionar una visión conjunta e integrada de la realidad. Esta visión integrada se logra con el *análisis multivariable*, formado por un conjunto de técnicas que resumen y sintetizan grandes conjuntos de datos buscando mejorar el conocimiento de la realidad.

Frente al *análisis bivariable*, el *multivariable* consigue una mayor economía en el almacenamiento de los datos, mayor consistencia en la inferencia estadística, desarrollo de conceptos teóricos más adecuados, y una mayor precisión y perspectiva conceptual.

El *análisis multivariable* supera ampliamente los objetivos de este texto, recomendando a los interesados la lectura de otros libros (por ejemplo Abascal y Grande 2005, Cea D'Ancona 2002, Díaz de Rada 2002, Luque 2000).

14. Se explicará en el próximo capítulo, apartado 4.

6. Anexo 1: Cuestionario utilizado como ejemplo

Número: 7____

Buenas tardes. En la asignatura *Introducción a la Investigación Social* estamos realizando una pequeña investigación para conocer el *ámbito social* del estudiante de Sociología. Para ello te presentamos un cuestionario sobre prácticas de ocio, y nos gustaría que lo respondieras sinceramente. Las respuestas de este cuestionario serán utilizadas para llevar a cabo las prácticas en esta asignatura durante todo el cuatrimestre. Por favor, CIRCULA la opción elegida

P.1 Refiriéndonos a lo que haces en un día de ocio, quisiéramos saber ¿cuál es la actividad que *más te gusta* hacer *fuera de casa* cuando dispones de tiempo libre? *[UNA RESPUESTA]*

(V01)

– BEBER, IR DE COPAS 01
– BAILAR 02
– HACER DEPORTE 03
– IR DE EXCURSIÓN 04
– VIAJAR 05
– R AL CINE 06
– IR AL TEATRO 07
– IR A MUSEOS 08
– IR A CONCIERTOS 09
– LEER LIBROS 10
– LEER PERIÓDICOS 11
– LEER REVISTAS 12
– PRÁCTICAR ALGUNA AFICIÓN O HOBBY 13
– OTRAS (apuntar) _____ 14
– NINGUNA EN PARTICULAR 15

P.2 Y, ¿cuál es la actividad que más te gusta hacer cuando *estás en casa*? *[UNA RESPUESTA]*

(V02)

– BEBER 01
– BAILAR 02
– VER LA TELEVISIÓN 03
– MANEJAR EL ORDENADOR 04
– JUGAR CON VIDEOJUEGOS, PLAYSTATION, ETC 05

– DORMIR, DESCANSAR, NO HACER NADA 06
– TRABAJAR EN LAS TAREAS DEL HOGAR 07
– ESTUDIAR 08
– ESCUCHAR MÚSICA 09
– LEER LIBROS 10
– LEER PERIÓDICOS 11
– LEER REVISTAS 12
– PRÁCTICAR UNA AFICIÓN O HOBBY 13
– OTRAS (apuntar) _____ 14
– NINGUNA EN PARTICULAR 15

P.3 De las siguientes situaciones, ¿podría indicar las *dos* que *mejor definen* tu actividad en tu tiempo libre? *[DOS RESPUESTAS]*

(V03)(V04)

– PASARLO BIEN SIN HACER NADA 1
– HACER MUCHAS COSAS, ESTAR ACTIVO, IR DE UN LADO A OTRO 2
– DEDICARME A LAS PERSONAS MÁS QUERIDAS 3
– HACER COSAS DE MI TRABAJO-ESTUDIOS QUE TENGO PENDIENTES 4
– DESCANSAR, RECUPERAR FUERZAS 5
– ESTAR CON LA GENTE, CHARLAR, TRATAR A LOS AMIGOS 6
– ABURRIRME 7
– PENSAR, MEDITAR 8
– DEDICARME TRANQUILAMENTE A MIS COSAS, MIS AFICIONES, DEPORTES . . . 9
– OTRAS: (APUNTAR) _____ 10

P.4 Normalmente, ¿cuántas horas libres tienes a la semana para tu ocio o diversión? ____ (v05)

P.5 En relación a tus hábitos de lectura, ¿me podrías decir cuántos libros *relacionados con tus estudios* has leído desde octubre del año pasado?
Número de libros _____ (v06)

P.6 Y de éstos, ¿cuántos eran de "*lectura obligatoria*"?
Número de libros _____ (v07)

P.7 Desde el inicio de tus estudios universitarios, ¿qué asignaturas recuerdas que proponen libros de "*lectura obligatoria*"? (apuntar todas) Letras mayúsculas
(v50)(v51)(v52)(v53)(v54) (v55)(v56)(v57)(v58)

P.8 A excepción de los libros relacionados con los estudios, ¿me podrías decir, aproximadamente, cuantos libros has leído desde octubre del año pasado?
Número de libros _____ (v08)

P.9 Aproximadamente, ¿cuántos libros de todo tipo tienes en tu hogar?
_____ (v09)

P.10 Habitualmente, ¿cuántos ejemplares de *periódicos* de *información general* lees por término medio a la semana? _____ (v10)

P.11 Habitualmente, ¿cuántos ejemplares de *periódicos* de *deportes* lees por término medio a la semana? _____ (v11)

P.12 ¿Cuántas revistas has leído desde navidades? _____ (v12)

P.13 ¿Y qué tipo de revista ha sido la última que has leído? *[UNA RESPUESTA]*
(v13)

– INFORMACIÓN GENERAL 01
– CORAZÓN 02
– MODA 03
– DEPORTIVA 04
– ECONOMÍA 05
– INFORMACIÓN TELEVISIÓN 06
– PROFESIONAL 07
– DECORACIÓN 08
– ERÓTICAS 09
– MOTOR 10
– PASATIEMPOS 11
– CIENTÍFICAS (Muy interesante, etc.) 12
– VIAJES 13
– ORDENADORES/INFORMÁTICA 14
– MASCULINAS (Man, Men's Health, etc) 15
– FEMENINAS (Cosmopolitan, etc.) 16
– MUSICAL 17
– HUMORÍSTICA 18
– OTRAS ¿Apuntar cuáles?_____ 19
– NINGUNA 20
– TODAS 21

Terminados los hábitos de lectura, dedicaremos unas preguntas a la televisión:

P.14 En un día laborable, ¿cuanto tiempo pasas viendo la televisión? (en minutos)
Apuntar minutos _____ (v14)

P.15 ¿Y en el fin de semana?, considerando conjuntamente el sábado y el domingo
_____ minutos (v15)

P.16 ¿Tienes vídeo y/o DVD?

(v16)
– SI 1
– NO . . . . . 2 *[IR A PREGUNTA 17]*

[*SOLO PARA LOS QUE TIENEN VÍDEO Ó DVD*]

P.16a ¿Tú o alguna otra persona que viva contigo ha visto alguna película de vídeo o DVD durante las últimas cuatro semanas?
– SI. ¿Cuántas películas?_____ (v17)
– NO . . . . . 0 *[IR A PREG. 16d]*

P.16b ¿Que día de la semana era cuando viste la última película de vídeo o DVD?

(v18)
– SÁBADO 1
– DOMINGO 2
– OTRO DÍA FESTIVO 3
– OTRO DÍA DE LA SEMANA NO FESTIVO . . 4

P.16c Durante el tiempo que dedicaste a ver el vídeo o DVD, ¿estabas sólo o acompañado?

(v19)
– ESTABA SOLO 1
– ESTABA ACOMPAÑADO 2
– NO RECUERDA 3

P.16d ¿Tú o alguna otra persona que viva contigo ha grabado en vídeo algún programa de televisión en las últimas cuatro semanas?

(v20)

– SI 1
– NO 2

El siguiente bloque de preguntas está referido al ámbito de la informática:

P.17 ¿Hay algún ordenador personal en tu hogar?

(v21)

– SI 1
– NO . . . . . 2 [*IR A PREGUNTA 18]*

[*SOLO PARA LOS QUE TIENEN ORDENADOR EN EL HOGAR*]

P.17a ¿Qué periféricos o dispositivos tiene/n el/los ordenadores de tu hogar?

– IMPRESORA 1 (v22)
– MODEM 1 (v23)
– ALTAVOCES 1 (v24)
– CÁMARA PARA CHATEAR CON IMAGEN (*WebCam*) 1 (v25)
– LECTORA DE CD ROM 1 (v26)
– LECTORA DE DVD 1 (v27)
– GRABADORA DE CD 1 (v28)
– GRABADORA DE DVD 1 (v29)

P.17b ¿Tienes acceso a Internet desde tu hogar?

(v30)

– SI 1
– NO 2

P.17c ¿Qué tipo de software y aplicaciones tienes para tu ordenador en casa?

(v31)(v32)(v33)(v34)(v35)(v36)

– PROCESADORES DE TEXTO 1
– HOJAS DE CÁLCULO 2
– BASES DE DATOS 3
– JUEGOS Y DIVERSIONES 4
– ENCICLOPEDIAS Y DICCIONARIOS 5

– PROGRAMAS EDUCATIVOS Y JUEGOS PARA APRENDER 6
– PROGRAMAS PRÁCTICOS PARA GESTIONAR ASUNTOS 7
– CD'S CULTURALES, MÚSICA CLÁSICA, ETC. 8
– OTROS (apuntar) _____ 9
– NINGUNO 10
– NO SABE 11

[*PARA TODOS*]

P.18 Tengas o no ordenador en el hogar. ¿Con que frecuencia utilizas...?

	TODOS O CASI TODOS LOS DÍAS	DOS O TRES VECES A LA SEMANA	UNA VEZ A LA SEMANA	MENOS DE UNA VEZ A LA SEMANA	NUNCA O CASI NUNCA	
Un ordenador, un PC	1	2	3	4	5	(v37)
Una conexión a Internet	1	2	3	4	5	(v38)
Correo electrónico	1	2	3	4	5	(v39)
Procesadores de texto	1	2	3	4	5	(v40)
Hojas de cálculo	1	2	3	4	5	(v41)
Bases de datos	1	2	3	4	5	(v42)

Para finalizar, unas preguntas sobre tus rasgos sociodemográficos:

P.19 ¿Con quién vives?

(v43)

– CON MIS PADRES 1
– CON AMIGOS EN UN PISO COMPARTIDO 2
– CON MI PAREJA 3
– OTRAS SITUACIONES (apuntar) _____

P.20 ¿Que titulación estás estudiando? _____ (v44)

P.21 ¿En que centro?

(v45)

– UNIVERSIDAD PÚBLICA DE NAVARRA-UPNA 1
– UNIVERSIDAD DE NAVARRA 2
– UNIVERSIDAD NACIONAL DE EDUCACIÓN A DISTANCIA-UNED 3
– OTROS CENTROS (apuntar) _____

P.22 ¿En que curso estás? _____ (v46)

P.23 ¿Cuántos hermanos tienes? _____ (v47)

[*SOLO PARA LOS QUE TIENEN HERMANOS*]

P.23a Considerando la edad, ¿qué lugar ocupas tú dentro de tus hermanos? (v48)

– SOY EL MAYOR 1
– SOY EL SEGUNDO MAYOR 2
– TENGO UNA EDAD INTERMEDIA ENTRE TODOS ELLOS 3
– SOY EL SEGUNDO MÁS JÓVEN 4
– SOY EL MAS JÓVEN 5

P.24 Para finalizar, no olvides apuntar tu sexo (v49)

– VARÓN 1
– MUJER 2

MUCHAS GRACIAS POR TU COLABORACIÓN

PARTE II

RECOGIDA Y DEPURACIÓN DE LA INFORMACIÓN

Capítulo III

Elaboración de un archivo de datos

1. Objetivos didácticos del capítulo

Tras la exposición del proceso de investigación en toda su amplitud llega el momento de recoger las opiniones expresadas por los entrevistados, proceso que se llevará a cabo en el presente capítulo. Comienza con una exposición de cómo deben prepararse los datos para posibilitar un correcto procesamiento de éstos utilizando el programa estadístico elegido. A continuación se procede con la conversión de "opiniones sociales" en un archivo de datos guardados en un formato magnético. El proceso es complicado puesto que es necesario cumplir una serie de etapas: definición de variables, entrada, y modificación de datos. Sin embargo, no será posible utilizar estos datos hasta haber realizado diferentes estrategias de *depuración* de la *información* recogida, aspecto que será explicado en el capítulo VI.

2. Aspectos previos a la introducción de datos: el formato de los datos

Expuestos los preliminares del análisis de datos (capítulo II), y una vez respondidos los cuestionarios utilizados como ejemplo, llega el momento de aplicar lo aprendido elaborando un archivo de datos donde trabajar con la información recopilada hasta el momento. Buscando explicar el proceso de la forma más sencilla posible, y considerando que la mejor forma de aprender a investigar es investigando, la elaboración del archivo de datos se va a realizar considerando un ejemplo, concretamente el cuestionario que se encuentra en el apartado 6 del segundo capítulo. Se trata de un cuestionario sobre prácticas de ocio, nivel de lectura y dominio de equipos informáticos respondido por la mayor parte de los estudiantes de la Licenciatura en Sociología de la Universidad Pública de Navarra desde el curso 2002/03. Es preciso indicar que el cuestionario fue elaborado, fundamentalmente, con un criterio pedagógico.

En el segundo capítulo ya señalamos que un archivo de datos se compone, en su estructura más básica, de filas y columnas: cada fila recoge información de un caso y en cada columna se representa una variable. En el ejemplo mostrado en el segundo capítulo, cuadro 2.2 en la página 35, las filas son las personas que prestan trabajo voluntario, mientras que las columnas recogen información sobre el sexo de cada uno, las horas que trabaja a la semana, si tiene una jornada partida o continua, etc.

Una vez elaborado el cuestionario, o el instrumento utilizado para la recogida de información, es necesario contemplar con detalle todas las posibles respuestas que pueden obtenerse de cada pregunta. Esto genera que el interés del investigador se desvíe del *número de preguntas del cuestionario* al número de respuestas –variables– que generan estas preguntas. Las *preguntas* "son la expresión manifiesta mediante la cual se recoge una determinada información" y están referidas a la estructura formal del cuestionario (Azofra, 1999: 9); mientras que las distintas informaciones incluidas en el cuestionario reciben el nombre de *variables*. Las variables, podríamos decir, son cada una de las *respuestas* o *informaciones* que se consignan en un cuestionario. En ocasiones las preguntas y las variables coinciden, pero la mayor parte de las veces no es así: una pregunta puede incluir más de una variable cuando se trata de preguntas de batería que presentan una serie de temas, o cuando no se limita el número de respuestas. Otras veces las variables no requieren de una pregunta, como es el caso del municipio donde se ha realizado el cuestionario, el sexo del entrevistado (en entrevistas personales), etc.

Tomaremos como ejemplo el cuestionario ya citado, situado en el apartado 6 del capítulo II, en el que puede apreciarse como el número de preguntas (19) es muy inferior al número de respuestas, al numero de variables utilizadas (58). Esta es la razón por la que cuando se comienza con el análisis de los cuestionarios se pierde el interés por las preguntas para centrarlo en las respuestas-variables. Por ello los cuestionarios suelen tener un código que acompaña a cada una de las posibles respuestas a fin de identificar cada variable del cuestionario (ver números entre paréntesis a la derecha de las preguntas en el cuestionario incluido en el apartado 2.6 del capítulo II). Debemos señalar, no obstante, que hay investigadores que utilizan el número de la pregunta (por ejemplo P2) y les añaden letras cuando tienen más de una respuesta (por ejemplo P2a, P2b, etc), otros prefieren colocar grupos de letras que resuman cada pregunta, etc. Más adelante dedicaremos algunas líneas a exponer los requisitos que deben cumplir los nombres de las variables.

En este trabajo se ha optado por definir cada posible respuesta con el número situado en el margen derecho de cada pregunta, número que estará acompañado de una "v". La decisión de utilizar esta opción está se fundamenta en que el SPSS muestra –en los menús de selección de variables– las variables por orden alfabético, o en su defecto con un orden numérico ascendente, de modo que definirlas así

permite localizar rápidamente cada una: la v05 estará antes que la v06, ésta antes que la v07, etc..

Aclarada la importancia de conocer de antemano el número de posibles respuestas obtenidas, llega el momento de tener en cuenta algunos aspectos relacionados con las características de las respuestas de cada pregunta. Es preciso recordar que el objetivo en este momento es construir un archivo de datos compuesto por "*n*" filas y "*m*" columnas, donde cada fila recoge las respuestas de un individuo a todas las preguntas y cada columna es la respuesta de todos los entrevistados a una pregunta concreta. Considerando conjuntamente el cuestionario referido y la figura 3.8 (ver página 62), la columna rotulada con el nombre v01 recoge todas las respuestas obtenidas por la pregunta "¿cuál es la actividad que más te gusta hacer fuera de casa cuando dispones de tiempo libre?", mientras que la fila 1 son las respuestas que un entrevistado ha dado a todo el cuestionario. Teniendo esto en cuenta, el número que aparece en la celdilla de la segunda columna y primera fila (valor 13) está indicando que a esta persona, lo que más le gusta hacer cuando tiene tiempo libre, es "practicar alguna afición o hobby".

Terminaremos este apartado exponiendo algunas ideas sobre la elección de los códigos para las opciones de respuesta; la codificación. Volvamos por un momento al cuadro 2.1 (página 34). Aquí se expone un ejemplo de codificación donde las dos variables se han codificado con letras, y el resto con números; con códigos alfanuméricos y numéricos respectivamente.

La decisión del tipo de códigos a utilizar corresponde totalmente al investigador, si bien recomendamos utilizar códigos numéricos porque todos los paquetes estadísticos aceptan estos códigos, mientras que algunos presentan problemas cuando se utilizan códigos alfanuméricos. El coste de introducción de datos es el mismo tanto si se introducen letras o números, pero la utilización de números hace nuestro trabajo más accesible a otros investigadores, además de permitir trabajar con una mayor diversidad de programas estadísticos. A estas ventajas hay que añadir que algunos paquetes estadísticos ofrecen la posibilidad de cambiar, en la pantalla de datos, los códigos numéricos por las etiquetas de los valores, siempre que se haya realizado una definición de las variables. En el SPSS esta opción se realiza seleccionando, del menú "*Ver*", la opción "*Etiquetas de valor*". Como se aprecia en la figura 3.9 (página 63), los números desaparecen y, en su lugar, se muestran las primeras palabras de las opciones elegidas por los entrevistados.

Por último, la experiencia recomienda introducir códigos en todas las celdillas, aún en las preguntas que no han sido respondidas, reservando por tanto un código para las no respuestas.

3. Consideraciones para a la creación de un archivo de datos: menú *vista de variables*

Antes de proceder con la construcción de un archivo de datos habrá que elaborar un archivo *diccionario* en el que quedan definidas todas las variables que formarán parte del archivo. Para ello, lógicamente, será preciso abrir el programa estadístico correspondiente, el SPSS 17.0 para Windows en nuestro caso.

Existen diversas formas de iniciar el paquete estadístico que utilizaremos en este texto, si bien la única forma para iniciar la *primera* sesión de trabajo[14] consiste en hacer un clic del ratón en el recuadro *Inicio* situado en la esquina inferior izquierda de la pantalla, para seleccionar a continuación la opción *Programas*[15], la carpeta *SPSS inc*, posteriormente *Statistics 17.0* y, por último, *SPSS Statistics 17.0*. Es posible que aparezca una pantalla que pregunta al usuario si desea trabajar con SPSS ejecutando el tutorial, tal y como se muestra en la figura 4.1, si bien no aparece siempre ya que depende de la configuración del programa. Cerraremos esta pantalla pulsando para ello el botón *Cancelar* –situado en la parte inferior de la pantalla– con el fin de visualizar el *Editor de datos* de SPSS; similar a una *hoja de cálculo* dividida en filas y columnas vacías (figura 3.1).

Téngase en cuenta que el programa siempre se inicia con un nuevo archivo de datos, como se muestra en la figura 3.1. En el caso que el lector se encuentre con el programa SPSS ya iniciado será necesario *crear* un nuevo archivo; seleccionando, del menú principal, las opciones *Archivo⇒Nuevo⇒Datos*. Se proceda de una u otra forma el resultado será la figura 3.1: Editor de datos de SPSS, vista de datos.

Considerando que nuestro primer objetivo es la definición de variables, será necesario pulsar la solapa *vista de variables* situada en la esquina inferior izquierda. También es posible acceder a la vista de variables haciendo doble clic en la parte superior de la ventana de datos, sobre el área sombreada donde se encuentra el nombre de la variable. Cualquiera de estas opciones presenta el editor de datos SPSS en la *vista de variables*, como se muestra en la figura 3.2.

El proceso de definición de una variable sigue un proceso que comienza con la elección del nombre, la definición del tipo de datos que la forman, la anchura de los datos (valor entero y decimal), las etiquetas de la variable, categorías de respuesta, valo-

14. Cuando existen archivos en SPSS es posible iniciar la sesión de trabajo haciendo doble clic sobre éstos, al igual que el resto de aplicaciones que funcionan en entorno Windows. Ahora bien, en el momento presente todavía no se ha utilizado el SPSS, de modo que suponemos que el lector no dispone de archivos SPSS para iniciar el programa de otra forma.
15. O *Todos los programas*, según la versión de Windows instalada en el equipo.

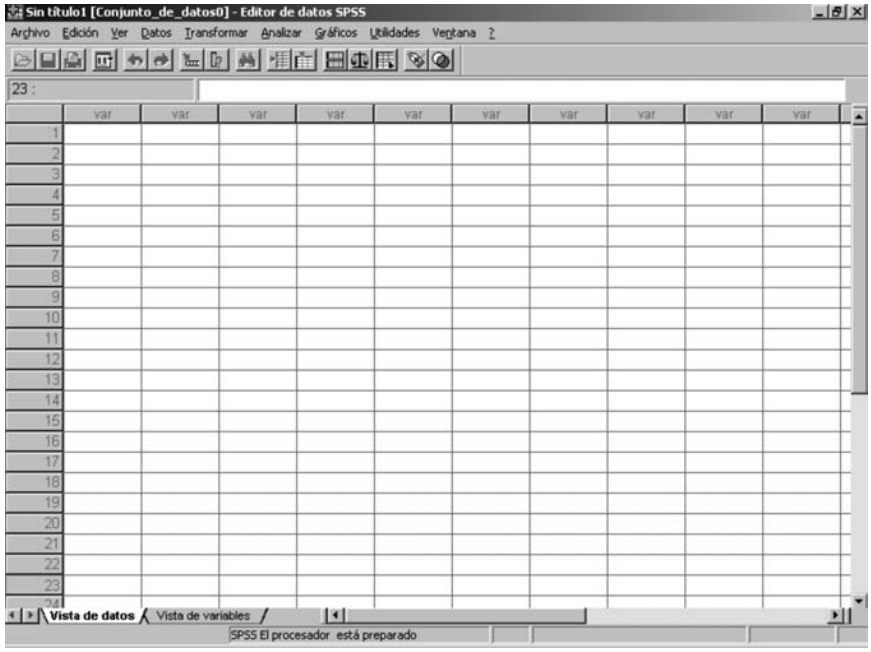


Figura 3.1. Menú principal: Editor de datos SPSS, vista de datos.

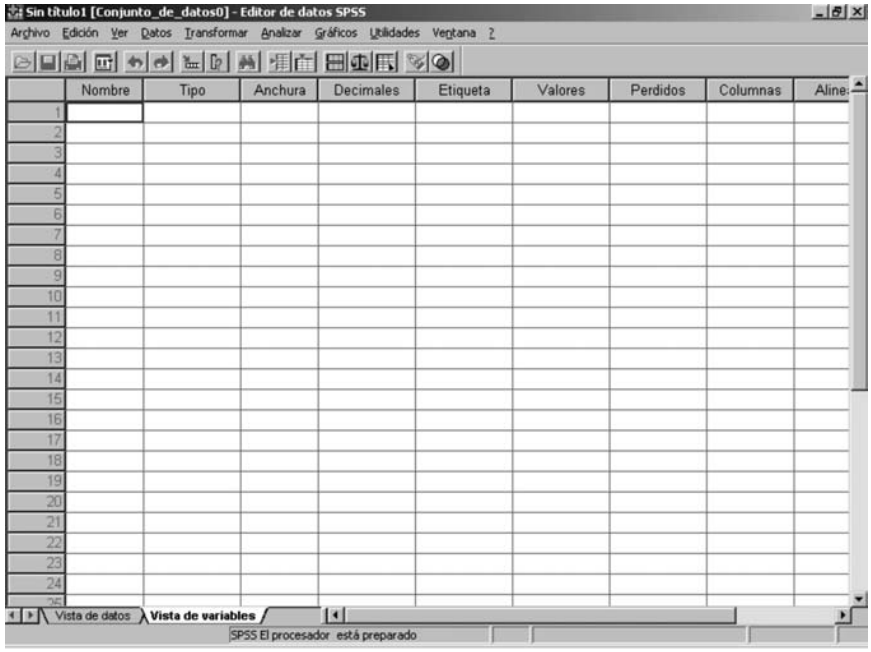


Figura 3.2. Menú principal: Editor de datos SPSS, vista de variables.

res perdidos, formato de columna, alineación y escala; aspectos que serán expuestos con detalle en las siguiente páginas. En la *vista de variables* cada una de las filas representa las variables de la investigación, y en las columnas aparecen diversos epígrafes para definir el *nombre* de la variable, *tipo*, *anchura*, *etiqueta*, etc. Analicemos en detalle cada uno de sus componentes:

- Nombre. Puede utilizarse cualquier nombre, siempre que cumpla los siguientes requisitos: como máximo deben tener una longitud de unos 64 caracteres y el primero de ellos siempre será una letra, no se admiten espacios en blancos en el nombre de una variable, tampoco asteriscos, ni interrogantes, ni exclamaciones, ni otros signos de puntuación. Otra de las restricciones es que no puede haber dos nombres de variables repetidos, y que no es posible utilizar las siguiente palabras por ser *operaciones lógicas* del SPSS: "ADD", "ALL", "AND", "ANY", "BY", "EQ", "GE", "GT", "LE", "LT", "NE", "NOT", "OR", "TO", "WIDT" y "WITH".

  Tampoco hay problema si se olvida alguna de estas restricciones puesto que al pulsar Enter (o el tabulador para desplazarse al siguiente elemento) aparecerá un mensaje diciendo "Nombre de variable no válido".

  Nuestra experiencia investigadora nos lleva a recomendar la utilización de nombres cortos; por ejemplo v01, v02, v03... por las razones apuntadas en el apartado 2.

  Cuando no se escribe nada en esta celdilla, o si se comienza a escribir en una línea distinta a la primera, el SPSS presenta automáticamente una variable llamada Var00001. Bastará con escribir sobre ella para que automáticamente este nombre se cambie por el que estamos escribiendo.
- Tipo. Definido el nombre de la variable es preciso indicar el tipo de datos que contiene. El SPSS permite utilizar ocho tipos de variables, aunque por defecto definirá las variables como numéricas con una anchura de 8 y dos decimales (ver figura 3.3). Veamos los elementos definitorios de cada tipo de variable:
  1. *Numérica*: admiten cualquier valor numérico, tanto positivo como negativo, teniendo una longitud máxima de 40 caracteres y admitiendo 16 decimales.
  2. *Coma:* Son variables numéricas que utilizan la coma para indicar los miles y el punto para separar la parte entera.
  3. *Punto*: como en el caso anterior, pero utilizando el punto para separar los miles y la coma para indicar decimales. Recomendamos definir de esta forma la variable donde se recogen el número de libros en el hogar (pregunta 9, variable v09).
  4. *Notación científica*: admite todos los valores numéricos más las letras D y E.
  5. *Fechas*.

Figura 3.3. Tipo de variable.

6. *Dólar*: a cada valor numérico introducido se le añade el símbolo dólar.
7. *Moneda personalizada*.
8. *Cadena*: variables alfanuméricas que admiten números y letras. Al definir este tipo de variables únicamente hay que indicar la longitud máxima, puesto que no tiene decimales. En los cuadros de diálogo estas variables aparecen con un símbolo alfanumérico a la izquierda, como puede comprobarse en la figura 3.4.

Figura 3.4. Cuadro de diálogo *Frecuencias*, con variables cadena (v01 y v02) y numéricas.

En el proceso de introducción de datos en las variables alfanuméricas podemos optar por introducir determinadas palabras de cada categoría de respues-

ta o bien los códigos numéricos adscritos a ellas (ver "etiquetas" unas líneas más adelante). Se trata, cuando se introducen los datos de la pregunta 1 (por ejemplo) de meter en el ordenador la palabra "bailar" o el valor "2", que sería su código correspondiente.

Ambas opciones son válidas aunque nosotros creemos que trabajar con códigos numéricos tiene algunas ventajas: en primer lugar agiliza tremendamente el proceso de introducción de datos puesto que es más rápido teclear el "2" que la palabra *bailar*. La introducción de palabras genera más errores puesto el programa SPSS considera que "bailar" es distinto de "BAILAR" y distinto de "Bailar", de modo que a la hora de pedir las frecuencias tendremos tres "bailar". Esto sin contar la posibilidad de cometer un error al escribir esta palabra. Otro problema es la necesidad de ampliar el formato de columnas de ciertas variables por si se introducen palabras de 10 y más caracteres.

Estas razones nos llevan a considerar totalmente desaconsejable introducir en el archivo de datos el texto de cada categoría de respuesta. Además, si se realiza una definición de las etiquetas de cada variable, siempre será posible ver las palabras de cada opción seleccionando –del menú *Ver*– la opción *Etiquetas de valor* (lo veremos en la figura 3.9).

Es preciso apuntar que las más utilizadas en la investigación con encuesta son las variables *Numéricas* y *Cadena*, si bien nos ha parecido interesante nombrar todas por las distintas utilizaciones que cada usuario pueden hacer del programa. Conviene señalar también que es posible cambiar la definición de una variable en cualquier momento de la investigación, aunque en numerosas ocasiones esto genera algunas pérdidas de información. Así, por ejemplo, al cambiar una variable numérica en nominal son eliminadas las etiquetas de las distintas opciones de respuesta.

Antes de terminar el proceso de elección del *tipo* de variable debe quedar claro que su definición tendrá implicaciones decisivas en el análisis de datos, en la medida que el programa restringe la utilización de determinados procedimientos estadísticos según el tipo de variable. Así, por ejemplo, las variables definidas como cadena no aparecen en el procedimiento "Estadísticos descriptivos", algo lógico por la imposibilidad de calcular los estadísticos en este tipo de variables. Algo similar ocurre en el procedimiento *Frecuencias* (que será explicado con detalle en el apartado 2 del capítulo VII): si bien el programa permite acceder al submenú *Estadísticos* aún con variables cadena, sin embargo no proporciona los estadísticos solicitados. En el próximo apartado se explicarán con detalle las implicaciones de cada tipo de variable en el análisis de datos.

Por motivos didácticos, en los ejemplos planteados hemos definido todas las variables como numéricas; algo que permitirá "equivocarnos" en el análisis de datos, así como verificar y corregir tales errores.

- Anchura. Ancho de columnas en el editor de datos, número de dígitos que necesita el editor para introducir correctamente los valores de cada variable. El editor de datos del SPSS aparece dividido en filas y columnas, generando (por defecto) celdillas con una amplitud de 8 dígitos y dos decimales. Si la variable a definir es más amplia será necesario aumentar el tamaño colocando un valor superior en esta opción. Observando la figura 3.8 (página 62) podemos apreciar que v01 necesita de dos dígitos puesto que cada respuesta está codificada con dos valores (12, 13, 14, etc), situación similar a v02; de modo que ambas variables necesitan una amplitud de columna de 2. Sin embargo, v16 tiene en sus respuestas números de un dígito, por lo que precisan de un ancho de columna de 1. En v09 un entrevistado da un valor de 1.000.000, situación que lleva a modificar la amplitud hasta colocarla en el valor 7[16] (ver figura 3.7, página 61). Si el ancho de la columna se deja en un valor inferior este valor no podrá ser visualizado en el editor de datos, apareciendo puntos suspensivos en su lugar (ver ejemplo en la figura 3.8).
- Decimales. Número de decimales en variables numéricas (esta celdilla queda *desactivada* en las variables tipo cadena).
- Etiqueta. Etiquetas de identificación de la variable. Estas etiquetas aparecerán siempre unidas al nombre de cada variable. La etiqueta de la variable puede ser cualquier texto con una longitud máxima de 256 caracteres, si bien una etiqueta tan larga no ofrece un buen efecto visual en determinados procesos estadísticos. Por este motivo aconsejamos limitarla entre 20 y 40 caracteres.
- Valores. Además de las etiquetas de las variables, SPSS ofrece la posibilidad de unir una definición a una de las distintas opciones de respuesta. Pulsando en la parte derecha de la casilla *valores* aparece el cuadro de diálogo mostrado en la figura 3.5. Para etiquetar las categorías de respuesta se coloca, en primer lugar, el número de respuesta en la ventana situada a la derecha de la palabra *valor*, y posteriormente la etiqueta de la variable en el espacio señalado para este fin. Tras hacer clic en el botón *Añadir* ambos valores se desplazarán a la ventana inferior, permitiendo así etiquetar otro valor. Una vez que todas las respuestas tienen su correspondiente etiqueta, el botón *Aceptar* dará por finalizado este proceso y mostrará de nuevo el Editor de datos en la vista de variables. Si, por un error, se pulsa *Aceptar* antes de *Añadir* el programa advertirá que "se perderá cualquier operación de Añadir o Cambiar pendiente".

16. Al tratarse de una variable "punto", es necesario ampliarla hasta 9: siete cifras, más los dos puntos de separación de miles.

Figura 3.5. Etiquetas de valor.

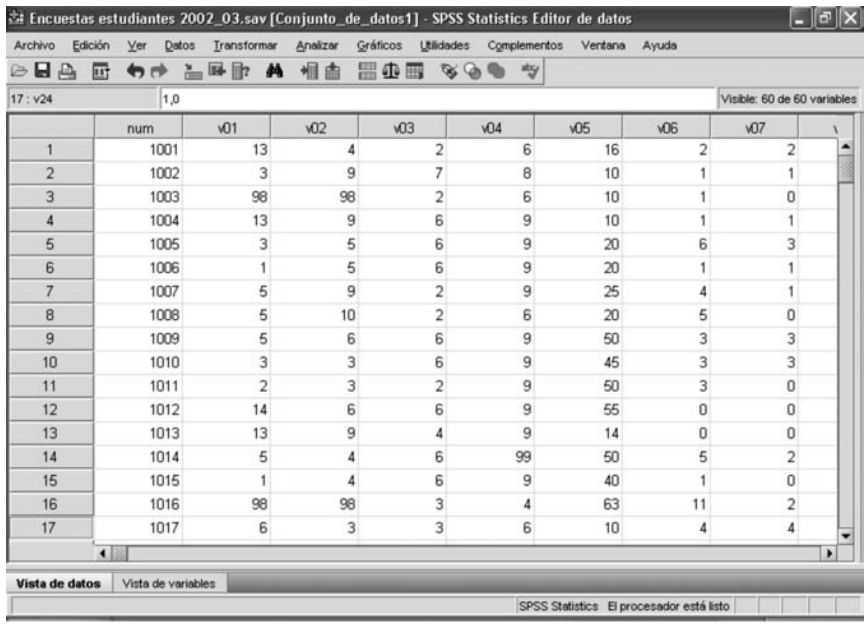
Respecto a la longitud máxima de las etiquetas de valor el programa SPSS permite hasta 60 caracteres, aunque aconsejamos utilizar nombres más cortos por la razón expuesta en el párrafo anterior sobre las etiquetas de variable.

- Valores perdidos. Valores que no son considerados en los análisis, pero que pueden utilizarse posteriormente. Anteriormente ya se han insistido en la conveniencia de introducir todos los valores, dando un valor determinado a la opción "no contesta". Se trata de evitar las celdillas vacías en la matriz de datos, puesto que con las celdillas vacías no se podrá realizar ninguna operación.
  Es posible que alguien se pregunte por las operaciones que pueden realizarse con los "no contesta": ¿qué tal si analizamos si éstas se distribuyen aleatoriamente o si presentan algún patrón determinado?; o quizás sea mejor analizar el perfil de los entrevistados que presentan más –y menos– no respuestas. Lo veremos más adelante.

Figura 3.6. Valores perdidos.

Hay varias formas para definir los valores perdidos: indicando los distintos valores en la segunda opción de la figura 3.6 (valores perdidos discretos); seleccionando el rango de valores perdidos, y utilizando ambas estrategias: rango más un valor perdido discreto.

- Columnas. Amplitud *total* de las columnas, que siempre debe ser superior a la anchura de la variable (referida 4 puntos más arriba). La diferencia entre ambas es que *Anchura* se refiere al espacio dentro de la ventana de datos, a la presentación de valores en el Editor, mientras que *Columnas* afecta a toda la variable, incluido su nombre. De este modo la amplitud de las *Columnas*, al estar referido a un ámbito mayor, siempre deberá ser superior a *Anchura*.
  Para modificar las columnas basta con introducir un valor en esta celdilla. Otra posibilidad es, en la *vista de datos* y situados en la línea del nombre de la variable[17] realizar la modificación colocando el ratón sobre el borde derecho de la celdilla, seguidamente pulsar el botón izquierdo del ratón y arrastrarlo hacia la derecha hasta conseguir la amplitud deseada. Lógicamente, este procedimiento únicamente puede realizarse sobre variables que han sido definidas previamente.



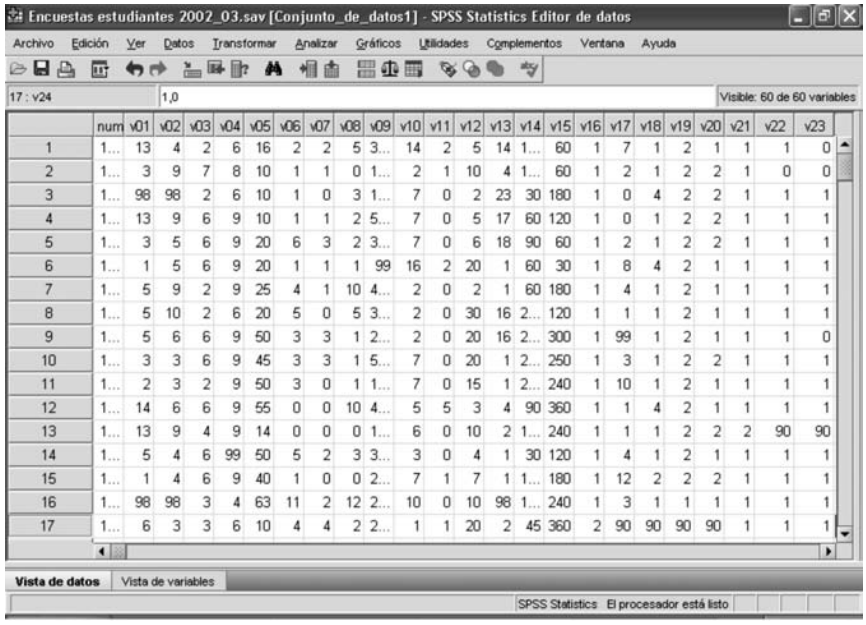
	num	v01	v02	v03	v04	v05	v06	v07
1	1001	13	4	2	6	16	2	2
2	1002	3	9	7	8	10	1	1
3	1003	98	98	2	6	10	1	0
4	1004	13	9	6	9	10	1	1
5	1005	3	5	6	9	20	6	3
6	1006	1	5	6	9	20	1	1
7	1007	5	9	2	9	25	4	1
8	1008	5	10	2	6	20	5	0
9	1009	5	6	6	9	50	3	3
10	1010	3	3	6	9	45	3	3
11	1011	2	3	2	9	50	3	0
12	1012	14	6	6	9	55	0	0
13	1013	13	9	4	9	14	0	0
14	1014	5	4	6	99	50	5	2
15	1015	1	4	6	9	40	1	0
16	1016	98	98	3	4	63	11	2
17	1017	6	3	3	6	10	4	4

Figura 3.7. Editor de datos SPSS: anchura de columna 8.

17. Línea sombreada en la parte superior de la vista de datos.

Decíamos que la amplitud por defecto es de 8 caracteres, pero ¿qué sucede si las variables utilizadas tienen una amplitud menor? En las versiones Windows del SPSS el Editor de datos aparece por defecto dividido en columnas con una amplitud de 8 caracteres, de modo que –aunque no es necesario modificarla siempre que las variables tengan menos de 8 caracteres– recomendamos hacerlo para tener una mayor *visibilidad* en el Editor de datos. Dicho de otra forma, es posible contemplar más variables de un vistazo; y como ejemplo basta con comparar el número de variables a la vista de la figura 3.7 y en la figura 4.2 (página 107). En la primera se han dejado todas las variables con la amplitud por defecto (8 caracteres), mientras que en la segunda se ha ajustado el ancho de columna al número de caracteres de la variable.

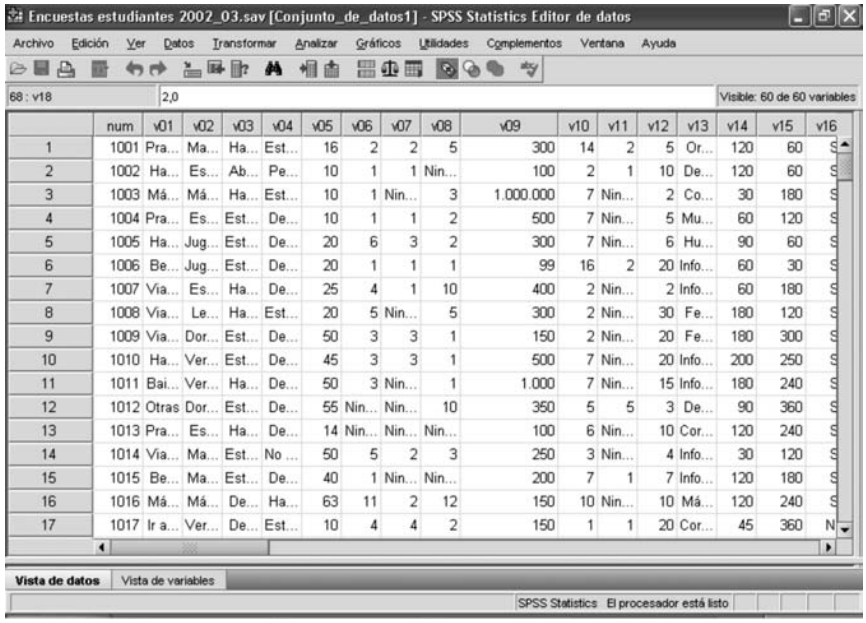
Unas líneas más atrás señalamos que la amplitud de las *Columnas*, al estar referido a un ámbito mayor, debe ser superior a la opción *Anchura*. ¿Qué ocurrirá si se coloca una menor amplitud, por ejemplo 2? En primer lugar que automáticamente se reduce la *Anchura* de cada variable, apareciendo puntos suspensivos cuando el ancho real de un valor es mayor que el ancho de la columna (esto sucede, en la figura 3.8, en la mayor parte de los valores de v09). Por otro lado, tan sólo serían visibles los primeros dígitos del nombre de cada variable, tal y como se puede apreciar en la figura 3.8.



	num	v01	v02	v03	v04	v05	v06	v07	v08	v09	v10	v11	v12	v13	v14	v15	v16	v17	v18	v19	v20	v21	v22	v23
1	1...	13	4	2	6	16	2	2	5	3...	14	2	5	14	1...	60	1	7	1	2	1	1	1	0
2	1...	3	9	7	8	10	1	1	0	1...	2	1	10	4	1...	60	1	2	1	2	2	1	0	0
3	1...	98	98	2	6	10	1	0	3	1...	7	0	2	23	30	180	1	0	4	2	2	1	1	1
4	1...	13	9	6	9	10	1	1	2	5...	7	0	5	17	60	120	1	0	1	2	2	1	1	1
5	1...	3	5	6	9	20	6	3	2	3...	7	0	6	18	90	60	1	2	1	2	2	1	1	1
6	1...	1	5	6	9	20	1	1	1	99	16	2	20	1	60	30	1	8	4	2	1	1	1	1
7	1...	5	9	2	9	25	4	1	10	4...	2	0	2	1	60	180	1	4	1	2	1	1	1	1
8	1...	5	10	2	6	20	5	0	5	3...	2	0	30	16	2...	120	1	1	1	2	1	1	1	1
9	1...	5	6	6	9	50	3	3	1	2...	2	0	20	16	2...	300	1	99	1	2	1	1	1	0
10	1...	3	3	6	9	45	3	3	1	5...	7	0	20	1	2...	250	1	3	1	2	2	1	1	1
11	1...	2	3	2	9	50	3	0	1	1...	7	0	15	1	2...	240	1	10	1	2	1	1	1	1
12	1...	14	6	6	9	55	0	0	10	4...	5	5	3	4	90	360	1	1	4	2	1	1	1	1
13	1...	13	9	4	9	14	0	0	0	1...	6	0	10	2	1...	240	1	1	1	2	2	2	90	90
14	1...	5	4	6	99	50	5	2	3	3...	3	0	4	1	30	120	1	4	1	2	1	1	1	1
15	1...	1	4	6	9	40	1	0	0	2...	7	1	7	1	1...	180	1	12	2	2	2	1	1	1
16	1...	98	98	3	4	63	11	2	12	2...	10	0	10	98	1...	240	1	3	1	1	1	1	1	1
17	1...	6	3	3	6	10	4	4	2	2...	1	1	20	2	45	360	2	90	90	90	90	1	1	1

Figura 3.8. Editor de datos SPSS: anchura de columna 2.

Por último, conviene señalar que reducir la amplitud de estas columnas plantea problemas a la hora de ver la etiqueta de las categorías de la variable cuando se elige la opción *Ver*⇒etiquetas de valor"; tal y como puede apreciarse con nitidez en la figura 3.9. Pese a esta situación, la experiencia investigadora nos ha enseñado que es preferible modificar el tamaño de las columnas para poder tener acceso visible a una mayor cantidad de variables, tal y como se mostrará en la figura 4.1.



	num	v01	v02	v03	v04	v05	v06	v07	v08	v09	v10	v11	v12	v13	v14	v15	v16
1	1001	Pra...	Ma...	Ha...	Est...	16	2	2	5	300	14	2	5	Or...	120	60	S
2	1002	Ha...	Es...	Ab...	Pe...	10	1	1	Nin...	100	2	1	10	De...	120	60	S
3	1003	Má...	Má...	Ha...	Est...	10	1	Nin...	3	1.000.000	7	Nin...	2	Co...	30	180	S
4	1004	Pra...	Es...	Est...	De...	10	1	1	2	500	7	Nin...	5	Mu...	60	120	S
5	1005	Ha...	Jug...	Est...	De...	20	6	3	2	300	7	Nin...	6	Hu...	90	60	S
6	1006	Be...	Jug...	Est...	De...	20	1	1	1	99	16	2	20	Info...	60	30	S
7	1007	Via...	Es...	Ha...	De...	25	4	1	10	400	2	Nin...	2	Info...	60	180	S
8	1008	Via...	Le...	Ha...	Est...	20	5	Nin...	5	300	2	Nin...	30	Fe...	180	120	S
9	1009	Via...	Dor...	Est...	De...	50	3	3	1	150	2	Nin...	20	Fe...	180	300	S
10	1010	Ha...	Ver...	Est...	De...	45	3	3	1	500	7	Nin...	20	Info...	200	250	S
11	1011	Bai...	Ver...	Ha...	De...	50	3	Nin...	1	1.000	7	Nin...	15	Info...	180	240	S
12	1012	Otras	Dor...	Est...	De...	55	Nin...	Nin...	10	350	5	5	3	De...	90	360	S
13	1013	Pra...	Es...	Ha...	De...	14	Nin...	Nin...	Nin...	100	6	Nin...	10	Cor...	120	240	S
14	1014	Via...	Ma...	Est...	No ...	50	5	2	3	250	3	Nin...	4	Info...	30	120	S
15	1015	Be...	Ma...	Est...	De...	40	1	Nin...	Nin...	200	7	1	7	Info...	120	180	S
16	1016	Má...	Má...	De...	Ha...	63	11	2	12	150	10	Nin...	10	Má...	120	240	S
17	1017	Ir a...	Ver...	De...	Est...	10	4	4	2	150	1	1	20	Cor...	45	360	N

Figura 3.9. Editor de datos SPSS: Ver⇒etiquetas de valor.

- Alineación. Posición (dentro de cada celdilla) donde se colocan los valores introducidos: a la izquierda de la celdilla, a la derecha y en el centro. Por defecto todos los valores aparecen alineados a la derecha, excepto cuando se trata de variables tipo *cadena* que son alineados a la izquierda.
- Medida. Escala de medida, con tres posibilidades: nominal, ordinal y escala. En variables numéricas aparecen las tres posibilidades, mientras que cuando se trata de variables alfanuméricas (tipo cadena) únicamente están disponibles las opciones nominal y ordinal. Cuando se define una variable como cadena el programa selecciona automáticamente la medida nominal. En el siguiente apartado se realiza una exposición detallada donde se explican las características de cada escala.

4. Tipos de variables considerando la escala de medida

Considerando la escala de medida (ó métrica de la variable) es posible distinguir entre variables categóricas o cualitativas, donde los elementos de variación tienen carácter no numérico, y variables numéricas o cuantitativas cuyos elementos tienen carácter numérico. Tras definir el proceso de medición como "la asignación de números a cosas y propiedades según ciertas reglas", a mediados del siglo XX Stevens formuló un concepto de medida para hechos sociales que sigue utilizándose hoy en día. Stevens (1946) elabora varios tipos de escalas según la regla utilizada en la asignación de códigos numéricos a las propiedades de los hechos sociales. Cada escala tiene distintas propiedades matemáticas que posibilitan utilizar determinadas pruebas estadísticas, de modo que la elección de cada escala condicionará la prueba estadística a utilizar.

Algunos hechos sociales pueden ser medidos utilizando una única escala, mientras que las propiedades de otros hechos posibilitarán la utilización de cualquier tipo de escala. Dos criterios deben utilizarse cuando se va a elegir una determinada escala: a) si una variable puede ser medida de diferentes formas deberemos utilizar el nivel de medida que posibilite utilizar los test estadísticos más poderosos, y b) el criterio que debe guiar la utilización de cada una es conseguir medir un fenómeno con la máxima precisión.

Medida nominal

La escala nominal es la más simple y débil forma de medición puesto que únicamente permite decir que las categorías difieren unas de otras. Los números sirven tan sólo para identificar o catalogar los objetos o sucesos, no siendo posible establecer ninguna relación de orden entre las categorías: una categoría de una variable no es necesariamente "mayor o menor" que otra, simplemente es diferente.

Estas escalas son utilizadas para las formas más sencillas de medición, permitiendo únicamente clasificar e identificar fenómenos. Presentan como restricción no asignar a dos categorías o sucesos el mismo número, o a dos números una misma categoría. La información debe ser clasificada en categorías no numéricas, exhaustivas y mutuamente excluyentes.

Ejemplos de escalas nominales son la medición del estado civil, sexo, voto político, lugar de residencia, matriculas de coche, etc. Algunos de estos ejemplos (sexo, por ejemplo) únicamente admiten dos posibles respuestas (hombre ó mujer), y por ello son definidas como variables dicotómicas.

En cuanto a las pruebas estadísticas a utilizar, al ser la forma más débil de medición es también la que permite utilizar un menor número de estadísticos. La tendencia central podrá ser conocida utilizando la moda, mientras que la dispersión y la distribución de los valores podrá conocerse mediante distribuciones de frecuencias. No es posible utilizar la media ni la mediana puesto que requieren propiedades del sistema de medición no conseguidas por esa escala.

Medida ordinal

La escala ordinal representa el siguiente nivel de medición ya que además de catalogar e identificar los objetos permite el establecimiento de relaciones de orden entre las diferentes categorías, siendo posible definir entre los individuos relaciones de preferencia, de mayor o menor que... A las relaciones de diferencia propias de las escalas nominales las ordinales añaden una diferencia de grado.

El estatus socioeconómico es un ejemplo de escala ordinal. Considerado un criterio definitorio (por ejemplo los ingresos) todos los miembros de la clase alta lo poseerán en mayor medida que los de la clase media, y estos a su vez más que la clase baja. Otros ejemplos de escala ordinal son el nivel de estudios, escala de posicionamiento político, etc.

En cuanto a las restricciones, además de las propias de las escalas nominales la medición ordinal debe respetar las relaciones observadas en la asignación del sistema de medición, ordenando los números según su orden serial. Los atributos o variables expresados en escala ordinal pueden cuantificarse por rangos; asignando a cada categoría el número de orden que le corresponda en la ordenación de tales categorías (p.e. nivel de estudios de mayor a menor, etc.).

Según las distintas posibilidades de cuantificación de las categorías es posible diferenciar dos tipos de escalas ordinales: las escalas de *fijación arbitraria* que se caracterizan por una asignación arbitraria de valores numéricos a las categorías manteniendo la información del factor original (la diferencia entre categorías), mientras que la *fijación por rangos* asigna a cada categoría el número de orden correspondido en la ordenación de tales categorías. Supongamos la variable nivel de estudios, formada por tres categorías: básicos, medios y superiores. La *fijación arbitraria* asignaría los valores -1, 0 y 1 respectivamente, mientras que en la *fijación por rangos* estas categorías adoptarían los valores 1, 2 y 3 respectivamente.

La escala ordinal permite utilizar más estadísticos que la anterior. Para conocer la tendencia central pueden utilizarse la moda y la mediana, muy interesante este último porque no varía con los casos extremos ni por la existencia de individuos atípicos. La dispersión de los datos puede ser conocida con el rango y la desviación inter-

cuartílica, mientras que los percentiles y las distribuciones de frecuencias permitirán conocer su distribución.

A lo largo de la exposición se ha comprobado como las escalas ordinales cumplen todas las propiedades y restricciones de las nominales, además de tener algunas propiedades más. En el párrafo anterior se ha puesto de manifiesto que los estadísticos utilizados en las escalas ordinales sirven también para las nominales, aunque ello implicará no considerar las relaciones de orden: al utilizar las escalas ordinales con los estadísticos propios de las nominales todas las variables son consideradas como nominales, es decir, no se tienen en cuenta las relaciones de orden entre las categorías. Evidentemente no es posible realizar la operación inversa; utilizar los estadísticos de las escalas ordinales para escalas nominales.

Como se ha afirmado las escalas ordinales miden si un objeto tiene más o menos de una característica con respecto a otro objeto, pero no proporciona información sobre la mayor o menor cantidad de esta característica que poseen el objeto. Es decir, sabemos que tiene una característica en mayor grado que otro, pero no es posible cuantificar la diferencia. Este será el elemento diferenciador del siguiente tipo de medición.

***Medida* escala** (Conocida en la literatura como escala de intervalo[18]).

Además de las propiedades de la escala ordinal, la escala de intervalo permite cuantificar numéricamente la distancia existente entre dos observaciones cualquiera. Otra característica distintiva de las escalas de intervalo es la igualdad de los intervalos: en esta escala la diferencia en la ordenación de los intervalos es expresada en unidades que tienen el mismo valor absoluto y que permanecen constantes a lo largo de toda la escala. De este modo el investigador puede especificar el número de características que tiene una categoría respecto de la anterior; "cuanto mayor" es una categoría respecto a otra.

Un ejemplo ayudará a expresar este concepto. Supongamos una variable de un cuestionario como podría ser la edad en años. Con esta variable es posible cuantificar numéricamente la distancia (en años) entre dos individuos, así diremos que el entrevistado de 30 años tiene 10 años más que el entrevistado de 20. Los intervalos de la escala son iguales al estar dividida en años, expresando cada entrevistado el número de unidades que posee de esta característica (años). A medida que aumenta o dismi-

18. De hecho, cuando se solicita la ayuda en el SPSS (Menú *Ayuda* ⇒ *Temas* ⇒"Asignación del nivel de medida") aparece el siguiente mensaje: "Puede especificar el nivel de medida como Escala (datos numéricos de una escala de *intervalo* o de *razón*), Ordinal o Nominal. Los datos nominales y ordinales pueden ser de cadena (alfanuméricos) o numéricos".

nuye la escala la unidad de medida no varía: en las edades superiores la medición no se realiza en quinquenios, ni en las edades inferiores se realiza por meses, y un año implica la misma cantidad de meses vividos; cumpliendo así que la diferencia entre los intervalos se expresa en unidades que tienen el mismo valor absoluto y que permanecen constantes a lo largo de toda la escala.

Respecto a las pruebas estadísticas disponibles, la escala de intervalo (denominada por otros autores como variable continua o cuantitativa), permite la utilización de todos los estadísticos descriptivos excepto la media geométrica, la media armónica y el coeficiente de variación. Para conocer la tendencia central se utiliza la moda, mediana, media, y media truncada (menos sensible a la presencia de casos extremos), mientras que la dispersión puede conocerse con la desviación típica, el recorrido intercuartílico y el rango. La especificidad de esta escala requiere en ocasiones conocer la asimetría y la curtosis de los datos recogidos. Además de estos estadísticos descriptivos, las escalas de intervalo permiten utilizar todos los procedimientos estadísticos paramétricos, siempre que los datos cumplan los requisitos especificados por cada técnica estadística.

Por último, es posible agrupar los distintos valores de una escala de intervalo formando así una escala ordinal. Al categorizar la edad de los entrevistados en tres grupos (entre 19 y 30 años, entre 31 y 40, y más de 40) obtendríamos una escala ordinal, y ya no sería posible cuantificar la distancia existente entre los que tienen 40 y los que tienen 20 años. Tan sólo puede decirse que unos son más viejos que otros.

Al principio de este apartado indicamos que determinados hechos sociales pueden ser medidos utilizando cualquier tipo de escala; recomendando (en estas ocasiones) utilizar el nivel de medida que posibilite utilizar los test estadísticos más poderosos, y aquellos que consiguen medir un fenómeno con la máxima precisión. Veámoslo con un ejemplo, relativo al nivel de asistencia al cine. Este hecho puede ser medido utilizando diferentes cuestiones:

a) Preguntando si va al cine.
   Respuestas: si/no.
b) Preguntando con que frecuencia va al cine.
   Respuestas: Todos o casi todos los días. Dos veces a la semana. Una vez a la semana. Dos o tres veces al mes. Una vez al mes y menos. Nunca.
c) Preguntando el número de veces que va al cine cada mes.
   Respuesta: anotar número de veces.

Es evidente que la calidad de la información recogida por la escala C es muy superior a la escala B, y ambas miden la asistencia al cine mejor que la escala A. A esto nos referimos cuando decimos que cuando se mida un fenómeno hay que intentar hacerlo con el máximo nivel de precisión. Además, siempre será posible "reducir" –en un

segundo momento– el nivel de precisión convirtiendo, en este ejemplo, la escala C en la B: los que van al cine 2 veces al mes pueden ser codificados en la escala B como "dos o tres veces al mes", los que van 6 días a la semana como "todos o casi todos los días", etc. Incluso es posible convertir las posibilidades de respuesta de ambas opciones a la dicotomía si/no de la escala A[19]. No es posible realizar el proceso contrario, pese a que se han realizado algunos intentos para lograrlo (Villarejo Ramos et al, 1996: 465-470).

Del párrafo anterior se desprende que la escala de intervalo constituye la mejor forma de medición, y de hecho es la más utilizada en el cuestionario utilizado como ejemplo en este texto (se mostró en el apartado 6 del segundo capítulo). Sin embargo, esta escala es también la menos utilizada en la investigación social y de mercado por el carácter *cualitativo* de los fenómenos investigados. Obsérvese, por ejemplo, el número de variables de intervalo presentes en el cuestionario sobre *Vida Cotidiana* que se muestra en la carpeta *capítulo 7* de los *materiales complementarios* (web).

5. Creación de un archivo de datos: definición de variables

En este apartado se realiza una aplicación práctica de lo explicado en los apartados 3 y 4, elaborando el archivo de datos para el cuestionario del apartado 6 del segundo capítulo; ejemplo que será utilizado a lo largo de todo el texto.

El proceso comienza, como ya se ha explicado, definiendo un nuevo archivo de datos (*Archivo⇒Nuevo⇒Datos*), para pulsar a continuación la solapa "vista de variables" (o haciendo doble clic sobre una variable). Comenzaremos con el nombre de la primera variable, que será el que aparece en la derecha de las preguntas, justo encima de los códigos de respuesta. Tras colocar v01 en la primera celdilla pasamos a definir el *tipo de variable*. Las actividades fuera de casa que más gusta hacer cuando se dispone de tiempo libre (que así se denomina la primera variable) no presentan ninguna relación de orden, de modo que se trata de una escala nominal, por lo que definimos la variable tipo *numérica*, con una a*nchura* de 2, y sin *decimales*. La *etiqueta* de la variable es "actividad fuera de casa que más gusta hacer cuando se dispone de tiempo libre", y en los *valores* reproducimos cada una de las opciones de respuesta.

19. Obsérvese, por su sencillez, como se ha operado con la pregunta 16 del cuestionario mostrado en el apartado 6 del segundo capítulo. Los entrevistados que NO han visto películas de vídeo o DVD en el período señalado son codificados con el cero, y los que han visto películas indican el número de películas vistas. En este ejemplo la transformación en una variable dicotómica (si/no) es realmente sencilla: el cero equivale a la categoría "no", y el resto de números al "si". En el capítulo VIII (apartados 4 y 5) se explicará como efectuar esta tarea con el SPSS.

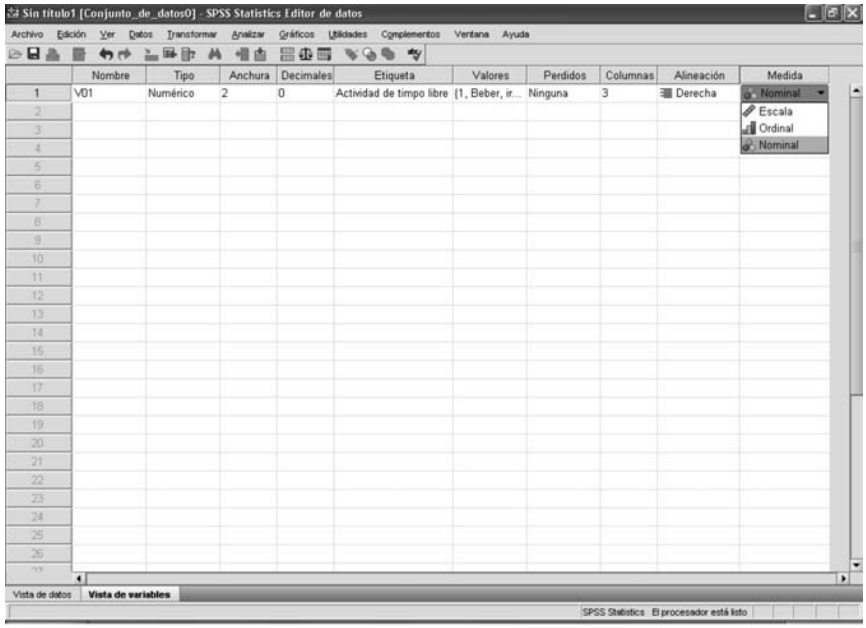


Figura 3.10. Editor de datos SPSS, Vista de variables, definición de la primera variable.

De momento no definimos los *valores perdidos*, dejando esta tarea para el proceso de depuración de la información recogida (capítulo VI).

La amplitud total de la *columna* es de 3, puesto que el nombre de la variable tiene tres dígitos ("v", "0" y "1"), dejamos la *alineación* por defecto, para terminar señalando que se trata de una medida *nominal*[20]. La definición de la medida de cada variable puede realizarse escribiendo en cada celdilla, o pulsando con el ratón en la parte derecha de la celdilla para ver las posibilidades que ofrece el programa. En la figura 3.10 se muestra el menú despegable de la *medida*: nominal, ordinal y escala. Seleccionado la primera de éstas, daremos por finalizado el proceso de definición de la variable v01[21].

Terminada la definición de la primera variable pasaremos a la segunda, después a la tercera,... y así hasta el final; recomendando dejar una variable para el número

20. Conviene señalar aquí que cuando una variable se define como numérica, automáticamente aparece seleccionada la opción "Escala".
21. Recordar que en la sección 3.3 se señaló que en los ejemplos planteados se definirán todas las variables como numéricas; algo que permitirá "equivocarnos" en el análisis de datos, con el fin de aprender a verificar y corregir tales errores.

de encuesta, definida como "num" en el *libro de códigos*. Considerando que el cuestionario (ver apartado 6 del capítulo II) comienza con la variable "Número", hubiera sido conveniente comenzar la definición del archivo con esta misma variable. Sin embargo la sencillez de esta variable, que recoge únicamente el número de encuesta y que no precisa definir etiquetas de valores, tipo de variable, etc. nos ha llevado a comenzar la exposición por v01, que presenta una definición más completa (y también más complicada). Explicada la definición de una variable más complicada, el lector deberá proceder con la elaboración del archivo de datos, recomendando que se comience con el número de encuesta.

La utilización de diversas herramientas presentes en el menú Edición pueden facilitar notablemente en el proceso de creación del archivo:

- Seleccionada una variable, mediante un clic de ratón en el margen izquierdo (parte sombreada) de la *vista de variables*, es posible copiar todas las características de ésta en otros lugares. Veámoslo con un ejemplo
  1. Seleccionamos V01, mediante un clic de ratón en el margen izquierdo (parte sombreada).
  2. Menú *Edición⇒Copiar*.
  3. Nos situamos en la segunda línea, y se selecciona con un clic de ratón en la parte sombreada del margen izquierdo.
  4. *Edición⇒Pegar*.
  5. Se copia toda la definición de la variable, excepto su nombre; que será automáticamente cambiando por el nombre que SPSS utiliza por defecto: Var00001.
  6. Cambiamos el nombre de la variable por v02. En el cuestionario del apartado 2.6 puede verse que ambas son muy similares, que se refieren a las actividades que se realizan en tiempo de ocio. La diferencia es que v01 recoge información del ocio *fuera de casa*, y v02 de las actividades de ocio en el *hogar*. Teniendo esto presente, bastaría con cambiar la *etiqueta* de la variable y los *valores* de respuesta.
  7. Situados en la celdilla *Etiqueta*, cambiamos "actividad, fuera de casa, que más gusta hacer cuando se dispone de tiempo libre" por "actividad, dentro de casa, que más gusta hacer cuando se dispone de tiempo libre".
  8. Situados en la celdilla *Valores*, cambiamos las categorías diferentes, o bien se copian directamente todas las categorías de v02.

  Este procedimiento facilitará tremendamente la definición de las variables v03 y v04, en la medida que se trata de dos variables con las mismas categorías de respuesta.

- Seleccionada una celdilla, es posible copiar estas características en otras variables. Supongamos que, después de realizar toda la definición del archivo deseamos alinear todos los valores en el margen izquierdo. El proceso sería de la siguiente forma:
  1. Situados en la celdilla *Alineación* de la primera variable, cambiamos Derecha por Izquierda.
  2. Menú *Edición⇒Copiar*.
  3. Nos situamos en la celdilla *Alineación* de la segunda variable.
  4. Menú *Edición⇒Pegar*.
  5. Nos situamos en la celdilla *Alineación* de la tercera variable.
  6. Menú *Edición⇒Pegar*.
  7. Y así sucesivamente hasta cambiar todas las variables deseadas.

  Este procedimiento facilitará tremendamente la definición de las variables de la pregunta 18 (v37–v42) puesto que todas presentan las mismas categorías de respuesta. De modo que bastaría con copiar la opción *Valores* y pegarlos en las siguientes variables.

  Aprovecharemos esta referencia a las variables de la pregunta 18, que como puede verse se trata de una escala ordinal, para explicar algunos aspectos que no quedan correctamente definidos en esta versión del SPSS. Anteriormente indicamos que al definir una variable como cadena no es posible solicitar estadísticos, si bien las escalas ordinales permiten utilizar la moda y la mediana para conocer la tendencia central, así como el rango y la desviación intercuartílica para vislumbrar la dispersión de los datos. De modo que cuando se utilizan variables medidas a nivel ordinal será necesario definirlas como numéricas, para seleccionar después la medida "ordinal". Solo de esta forma será posible solicitar los estadísticos.

Conviene recordar que es posible aumentar la rapidez del proceso explicado más arriba copiando y pegando determinados atributos de las variables ya definidas utilizando el botón secundario (derecho) del ratón (figura 3.11). Colocados en la *Vista de variables*, y sobre el atributo de la variable que se desea copiar, pulsando el botón derecho del ratón se despliega el menú contextual que aparece en la figura 3.11, en el que será necesario marcar la opción *Copiar*. Posteriormente, situados en el lugar de

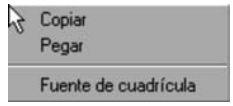


Figura 3.11. Menú contextual. Botón derecho del ratón sobre *Vista de variables*.

la nueva variable, bastará con pulsar el botón derecho del ratón y seleccionar la opción *Pegar*.

Dentro de la *Vista de variables*, y situados en la parte sombreada (margen izquierdo de la pantalla) pulsando el botón secundario del ratón aparece otro menú contextual que permite *Cortar* variables, *Copiar*, *Pegar*, *Borrar*, *Fijar columnas seleccionadas*, *Deshacer la fijación*, *Ordenar de forma ascendente* y de forma descendente; tal y como se muestra en la figura 3.12.

Figura 3.12. Menú contextual. Botón derecho del ratón sobre el nombre de la variable, dentro del *Vista de variables*.

La definición de variables no estará finalizada hasta que no se guarde este archivo con el menú *Archivo⇒Guardar*, o pulsando conjuntamente las teclas Control+G: el programa pedirá un nombre de archivo[22], y tras pulsar "*Guardar*" daremos por finalizado el proceso de definición de variables. En la figura 3.13 se muestra el editor de datos, en vista de variables, tras definir todas las variables del cuestionario.

Al finalizar la definición quizás no recordemos alguna información de determinadas variables, puesto que si el cuestionario es muy largo y este proceso ha sido realizado en varias sesiones se han podido olvidar determinados rasgos. En el menú *Utilidades⇒Variables* el programa muestra un cuadro de diálogo que proporciona información sobre cada una de las variables existentes (figura 3.14): nombre, etiqueta, tipo de variable, valores perdidos y etiquetas de valores. Este cuadro de diálogo –que normalmente se utiliza para comprobar que todas las variables están correctamente definidas– también permite un desplazamiento rápido al archivo de datos cuando es preciso realizar alguna corrección: seleccionada una variable, marcando el botón "Ir a" se accede a la posición de ésta en el Editor de datos.

Es posible enviar la información de la figura 3.14 al fichero de resultados seleccionando el menú *Archivo⇒Mostrar información del Archivo de datos⇒Archivo de trabajo*. Recomendamos imprimirla y leerla detalladamente para asegurarse que es igual que el *libro de códigos*.

22. A este nombre el SPSS le añadirá automáticamente la extensión ".sav".

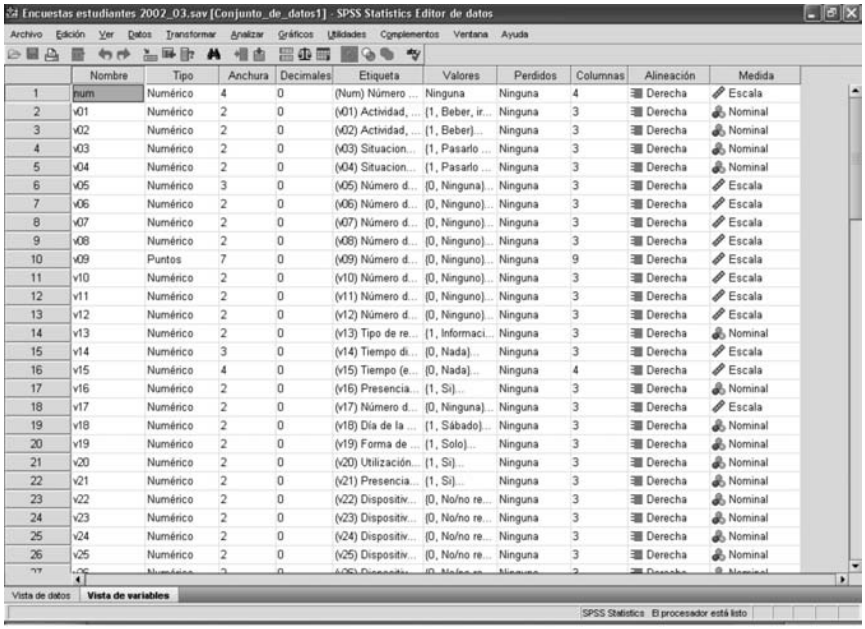


Figura 3.13. Menú principal: Editor de datos SPSS, *Vista de variables*.

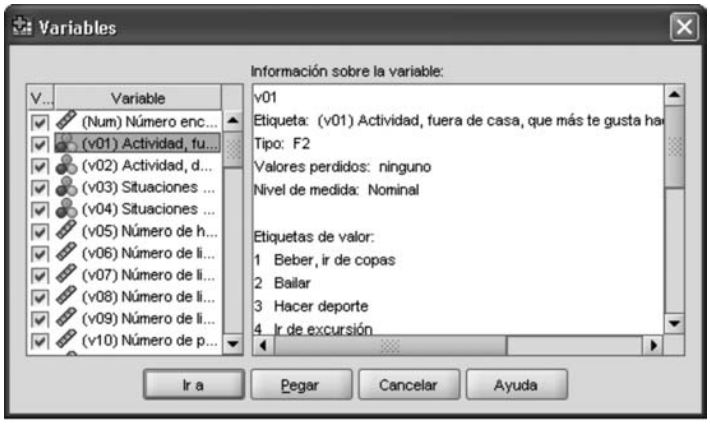


Figura 3.14. Cuadro de diálogo *Información sobre variables*.

6. Preparación de los datos: codificación de respuestas

En el segundo capítulo, dentro del apartado 2.2, se señalo que antes de proceder con la elaboración del archivo de datos hay que *trasladar* las respuestas del cuestionario a información analizable, almacenando esta información mediante una representación numérica o por medio de otros símbolos (proceso conocido con el nombre de *codificación*). Recordemos que el proceso de codificación se compone de dos tareas:

1. En un primer momento se introducen códigos en el cuestionario, o en el formato elegido para la recogida de la información, con el fin de identificar cada pregunta con un nombre determinado. En el apartado 2 de este capítulo ya señalamos que en el ejemplo que estamos utilizando cada una de las respuestas-variables han sido identificadas con un código numérico situado (entre paréntesis) a la derecha de cada pregunta del cuestionario.
2. En un segundo momento se proceder con la elaboración de un *libro de códigos* donde se recogen todas las categorías de respuestas. Volviendo al cuestionario del ejemplo se aprecia que "beber, ir de copas" está codificado con un 1, "bailar" con un 2, "hacer deporte" con un 3, etc.

Ahora bien, ¿qué código se adopta cuando alguien no responde una pregunta?; ¿o cuándo no debe responderla porque ha sido *filtrada* por otra pregunta anterior?; ¿o en las situaciones que el entrevistado procede a marcar dos respuestas en vez de una? Nada puede dejarse al azar, es necesario fijar una forma común de asignación de códigos para que todas las personas que participan en este investigación utilicen el mismo sistema de clasificación. En este momento es preciso adoptar ciertas *convenciones* que deben quedar muy claras para todos los miembros del equipo investigador, y para ello serán apuntadas en el *libro de códigos*. El *libro de códigos* utilizado para el cuestionario respondido en clase se adjunta en el apartado 9, y a en los siguientes párrafos se realizará una explicación sobre su proceso de elaboración.

Otro aspecto importante es que –como se ha señalado anteriormente– hay que rellenar todas las celdillas, todas las celdillas deben contener información. Es posible transformar un "0" o cualquier otro valor; pero es mucho más complicado realizar cualquier operación (aún la más elemental) con una celda vacía (con *nada*).

La primera variable, definida como *num*, recoge el número de encuesta. A continuación la pregunta 1 del cuestionario, cuyos valores quedan recogidas en la variable *v01*, y que ha sido etiquetada como *actividad –fuera de casa– que más gusta hacer cuando se dispone de tiempo libre*. Bajo esta definición se muestran las opciones de respuesta incluidas en la pregunta (ver cuestionario en la sección 2.6, y libro de códigos en la 3.9), codificadas de forma idéntica al cuestionario pero sin el cero a la izquier-

da que acompaña a los números menores de diez[23]. Otra diferencia entre el cuestionario y el libro de códigos es la aparición, en este último, de algunas categorías nuevas. Como puede apreciarse la primera pregunta del cuestionario presentaba 15 opciones de respuesta, mientras que en el libro de códigos aparecen algunas más. Las codificadas con los valores del 16 al 18 tienen su origen en la opción *otras (apuntar)*. Algunos entrevistados han respondido *otras* actividades distintas a las mostradas en la pregunta, apuntando a continuación el nombre de tales actividades (ver opción 14 en la pregunta 1 del cuestionario). Estas actividades deben ser codificadas, y en este ejemplo han adoptado el valor 16 para *quedar con amigos/as*, el 17 para *quedar con el novio/a*, y el 18 para *ir al monte*. Lógicamente, aquellos que responden la opción "otras" y que no señalan el nombre de la actividad no pueden ser recolocados en nuevas categorías, por lo que son codificados con el valor 14.

Respecto a las *nuevas* categorías colocadas al final de la pregunta, se ha utilizado el valor 99 cuando un entrevistado no responde a una pregunta (recordar que más atrás recomendamos que todas las celdillas tuvieran información), y el 98 cuando se elige más de una respuesta en preguntas de respuesta única (o cuando son elegidas más de dos en preguntas de dos respuestas). Explicaremos pormenorizadamente la razón de asignar un valor específico para esta situación: imaginemos que el primer entrevistado ha elegido, en la pregunta 1, las opciones *bailar* y *practicar alguna afición o hobby*. Como en el enunciado de la pregunta aparece la instrucción una respuesta, el archivo de datos se ha elaborado pensando en introducir una única respuesta en esta pregunta. En la situación en la que nos encontramos, ¿con qué valor codificaremos la respuesta dada?; o mejor aún, ¿qué valor introduciremos en el ordenador?:¿el 2 (bailar), o el 13 (practicar alguna afición o hobby?). Ninguno de éstos puesto la persona que introduce los datos no puede decidir la actividad que más le gusta hacer al entrevistado. En esta situación, en la que el cuestionario ha sido respondido erróneamente, conviene dejar un código para reflejar esta situación, indicando que la pregunta se ha respondido incorrectamente[24].

Antes de terminar con la explicación de esta pregunta conviene señalar que estas cinco *nuevas* categorías aparecen en negrilla en el libro de códigos, con el fin de diferenciarlas de las categorías incluidas en el cuestionario. Alguien podría pensar, ¿por

23. Se funciona de esta forma porque es más sencillo, y más rápido, introducir el valor "1" que el "01". De forma similar se ha actuado en la pregunta 2, pero no así en la pregunta 13 (v13) donde se dejan los valores originarios del cuestionario para que se comprenda hasta que punto eliminar el 0 antes de los números menores de diez supone un menor esfuerzo.
Sea cual sea el criterio adoptado para la codificación es necesario que todo el equipo de investigación tenga claro este criterio para que todos introduzcan los mismos valores.

24. En otro texto (Díaz de Rada 2001) se analiza en detalle cómo formular preguntas correctamente, al tiempo que se utilizan *indicadores* que muestran preguntas mal formuladas. Esta situación, una pregunta que no se responde adecuadamente, es una pregunta mal formulada.

qué no se colocan estos valores en el cuestionario? Al tratarse de un cuestionario autorrellenado no es recomendable colocar una categoría para la opción *no respuesta*. Además, siempre deberíamos utilizar determinados valores para identificar dos respuestas en preguntas de una respuesta, así como para las preguntas *filtradas*. De modo que el libro de códigos es un *compañero* inseparable del investigador a partir de este momento, debiendo estar accesible tanto en el proceso de introducción como en el análisis de datos.

Un repaso del resto de preguntas incluidas en el libro de cogidos muestra una situación similar en la pregunta segunda y en la tercera (variables v02, v03 y v04). Respecto a la cuarta pregunta (variable v05), "número de horas libres para ocio o diversión", no se han etiquetado los valores de respuesta puesto que se trata de una variable numérica, cuantitativa, una escala de intervalo. Ante el temor que alguien pudiera señalar que tiene 99 horas libres, se ha decidido *cambiar* el código no respuesta utilizado hasta ahora, sustituyéndolo por el valor 999. Aunque es difícil que alguien tenga 99 horas libres a la semana (14,14 horas libres al día), elegir el valor 999 para las no respuestas elimina la posibilidad de error[25]. Este mismo argumento explica el valor utilizado para codificar la no respuesta de la pregunta nueve[26] (v09, número de libros en el hogar), en la catorce y en la quince (v14 y v15), tiempo diario viendo la televisión).

La pregunta 16a (variable v17) muestra una situación no contemplada hasta el momento, en la medida que es respondida *únicamente* por aquellos que tienen vídeo o DVD en el hogar. A la hora de codificar estas respuestas hay que diferenciar los que no responden porque no desean (u olvidan hacerlo), de aquellos que no lo hacen porque no tiene lugar esta pregunta, es decir, porque no tienen vídeo o DVD. Así, además del valor 99 (para las no respuestas) hay que considerar otro valor para aquellos que no tienen vídeo, que no deben leer (ni responder) esta pregunta. Cuando no proceda la respuesta porque una determinada pregunta ha sido *filtrada* por otra, se colocará el valor 90. Esta misma situación sucede en la pregunta 16b, 16c y 16d (variables v18, v19 y v20).

La pregunta 17 (variable v21) no presenta ninguna complicación, pero sí la pregunta 17a (variables v22 a v29) que, al estar compuestas por distintas opciones, se conoce como una pregunta multirespuesta dicotómica. Es una pregunta que puede obtener –como máximo– ocho respuestas (proporcionadas por los entrevistados que tienen ordenadores con todos los dispositivos mostrados), de modo que es necesario dejar una variable para cada respuesta; en este caso de la v22 hasta la v29. Los dispositivos

25. Aunque alguien tuviera exactamente 99 horas libres a la semana, es más normal que señale "en torno a 100" (aproximadamente 100) que "exactamente 99".
26. La codificación de la pregunta 7 (variables v50–v58) se realizará más adelante.

presentes en los equipos de los entrevistados son codificados con el valor 1, y la ausencia de éstos con el 0, no con el 99 de *no responde*. Téngase en cuenta que en el cuestionario sólo aparece el valor 1.

La pregunta 17c (variables v31 a v36) es también una pregunta multirespuesta pero, a diferencia de la 17a, presenta un número limitado de respuestas. Aquí se pregunta sobre el tipo de software y/o aplicaciones que cada entrevistado tiene en su ordenador y, aunque hay 11 posibilidades de respuesta, únicamente hay espacio para seis respuestas (variables v31 a v36). Debe quedar claro que la pregunta 17b y la 17c son dos tipos de preguntas multirespuesta diferentes, y normalmente cada investigador muestra sus preferencias por utilizar una u otra, de modo que es extraño alternar ambos tipos de preguntas multirespuesta en un mismo cuestionario. Aquí se presentan ambas con un fin didáctico, para que los lectores vean las diferencias entre unas y otras, diferencias a nivel de diseño y a nivel de análisis[27]

Las diferencias entre ambas no sólo se refieren al número de respuestas, sino también al criterio de *codificación* y al *coste* en el procesado de cada una. En las preguntas multirespuesta categóricas (pregunta 17c) cada posibilidad de respuesta está codificada con un número diferente (p.e. procesador de texto con el 1, hojas de cálculo con el 2, bases de datos con el 3, etc.), mientras que en la multirespuesta dicotómica todas están codificadas con el mismo valor (el 1 en el caso del ejemplo).

En relación al coste, la pregunta multirespuesta dicotómica es más costosa puesto que –independientemente del número de respuestas– precisa introducir un valores en todas las posibles respuestas. En el caso de la pregunta 17a, por ejemplo, la respuesta de un entrevistado que dice no disponer de ninguno de esos dispositivos será "0" en v22, "0" en v23, "0" en v24... y así hasta v29. Es decir, disponga o no de estos dispositivos hay que introducir 8 valores por cada entrevistado. Por otro lado, en la pregunta 17c (multirespuesta categórica) únicamente hay que introducir 6 valores, dos menos que en la pregunta 17a. Quizás en este ejemplo no se aprecie la importancia entre introducir 8 o 6 valores (preguntas 17a y 17c respectivamente) pero, ¿qué ocurriría si tuviéramos 5.000 cuestionarios[28]? Sencillamente, utilizar una pregunta u otra implicaría la introducción de 10.000 dígitos más o menos. En el año 2009 la introducción de cada dígito se cobra a 0,0325 Euros (más IVA), de modo que optar por una pregunta u otra genera una diferencia de 32,5 Euros. Aquí hay que añadir el coste por lectura de datos, que aumenta a medida que el archivo es mayor.

27. En otro texto (Díaz de Rada, 2001) hemos dedicado más atención al diseño y utilización de estas preguntas.
28. Algo, por otro lado, frecuente. Téngase en cuenta, por ejemplo, que los estudios del Centro de Investigaciones Sociológicas realizan mensualmente 2.500 entrevistas.

Sin embargo no todo son ventajas en las preguntas multirespuesta categóricas. ¿Qué ocurre cuando, por ejemplo, alguien señala que tiene más de seis aplicaciones en su ordenador?, por seguir con el ejemplo de la pregunta 17c. En este caso no queda más remedio que ampliar el número de variables o, más sencillo, cambiar la codificación de la pregunta. En el presente ejemplo se ha optado por utilizar un valor (el 12) para aquellos que disponen de más de seis aplicaciones, y el valor 90 (no procede) en el resto de variables de esta pregunta[29]. Debe tenerse en cuenta que esta forma de proceder impedirá conocer las aplicaciones disponibles en los ordenadores de aquellos entrevistados que dan más de seis respuestas. Aquellos entrevistados que tienen más de 6 aplicaciones han sido codificados con el valor 12, y no se han anotado los valores de cada una de las aplicaciones. Digamos que es una *solución de urgencia* que puede empañar bastante los resultados de la investigación[30]. Por este motivo conviene utilizar estas preguntas únicamente cuando se limita el número de respuestas del entrevistado[31], o cuando se conoce –con bastante precisión– que no se va a sobrepasar un determinado número de respuestas[32]. En este ejemplo se ha utilizado este tipo de pregunta porque el objetivo es conocer el tipo de aplicaciones presentes en los ordenadores con menos aplicaciones; objetivo que se logra perfectamente utilizando esta pregunta.

La pregunta 18 (variables v37 a v42) es una pregunta de batería formada por seis variables, y no plantea dificultades puesto que no supone variaciones respecto de las normas expuestas con anterioridad. Lo mismo cabe decir de la pregunta 19 (variable v43), si bien aquí se han introducido nuevos códigos con otras posibles situaciones: valor 4 cuando vive con padre ó madre; 5 si lo hace con familia propia; 6 para hermanos; 7 con madre, madre y hermanos; 8 cuando vive con tíos y primos; 9 con madre y abuela; 10 cuando vive en una residencia de estudiantes o un colegio mayor; 11 si vive solo; y 12 cuando vive con otros familiares

La siguiente (pregunta número 20, variable v44) es una pregunta *abierta* en las que el propio entrevistado debe anotar la respuesta, en vez de elegirla de una serie de opciones (como en las preguntas anteriores). Al desconocer las posibles titulaciones de los amigos de los estudiantes de Sociología se ha creído conveniente dejar un espa-

29. Es decir, en vez de colocar uno a uno los valores de cada aplicación, se codifica el 12 en v31 cuando el entrevistado declara que tiene en su ordenador más de seis aplicaciones, y el valor 90 en el resto de variables.
30. Esta solución es adecuada cuando sea más importante el número de respuestas que la categoría elegida. De modo que solo es aplicable cuando el objetivo es analizar el software que aquellos entrevistados que usan poco el ordenador (aquellos que proporcionan menos de seis respuestas).
31. Ejemplos: ¿Cuáles son, a su juicio, los *tres* problemas principales que existen actualmente en España?; ¿cuáles son sus *tres* programas (de televisión) favoritos?, etc...
32. Por ejemplo dejar nueve respuestas para la pregunta ¿qué asignaturas proponen libros de lectura obligatoria?, como se hace en la pregunta 7 del cuestionario.

cio para que cada entrevistado escriba su titulación. Las respuestas obtenidas se codifican asignando el número correspondiente del libro de códigos: los entrevistados que estudian LADE obtienen un 1 en esa pregunta, un 2 los estudiantes de la Licenciatura en Economía, el 3 para los que estudian la Diplomatura en Ciencias empresariales, el 4 para los Licenciados en Sociología, etc. Las preguntas 22 y 23 (curso y número de hermanos) son también preguntas *abiertas*, si bien más restringidas en sus respuestas puesto que deben responderse con un número.

Una situación similar a la pregunta 20 presenta la pregunta 7 (variables v50 a v58), que recoge información sobre las asignaturas que proponen libros de lectura obligatoria y que exige del entrevistado que anote los nombres de las asignaturas (pregunta abierta). Con esta pregunta se buscaban dos objetivos: conocer el *número* de asignaturas que obligan a leer y después *identificar* tales asignaturas. Para llevar a cabo la codificación de esta pregunta se realizó un listado exhaustivo de todas las asignaturas del Plan de Estudios de Sociología, numerándolas después desde el valor 1 al 41, y éstos son los valores que han sido introducidos en el libro de códigos. Los entrevistados que afirmaron que *ninguna* asignatura obliga a leer fueron codificados con el 0, con el fin de diferenciarlos de aquellos que no respondieron la pregunta. Así, cuando el entrevistado señale que ninguna asignatura obliga a leer (que ninguna propone libros de lectura obligatoria) se colocará el "0" en la variable 50 y en el resto de variables el valor 90 ("no procede"), de forma similar a como se procedió en la pregunta 17c.

En los estudiantes de Sociología la codificación no reviste problemas pero, ¿cómo debe procederse con sus amigos, aquellos que estudian otras carreras? Ante la dificultad de anotar (y codificar) todas las asignaturas de este colectivo tan heterogéneo se ha optado por codificarlas con un código, concretamente el 51. Una situación similar se ha adoptado para recoger la información de las asignaturas que –no siendo específicas del Plan de Estudios de Sociología– han sido elegidas por los estudiantes de Sociología (nos referimos, lógicamente, a asignaturas de libre elección). En este caso se ha adoptado el código 50.

Esta codificación permite, en el caso de los estudiantes de Sociología conocer el número y el nombre de las asignaturas que proponen libros de lectura obligatoria. En el caso de los que no estudian Sociología tan sólo es posible conocer el número de asignaturas que obligan a leer, no el nombre de cada una. Al haber tanta variabilidad de carreras el nombre de las asignaturas carece de importancia, restringiendo la investigación al conocimiento del *número* de asignaturas que proponen libros de lectura obligatoria[33].

33. Esta situación, junto a la ocurrida en la pregunta 17c, nos lleva a insistir que tan importante como la redacción de las preguntas es el *uso* que haremos de ellas. Por este motivo cuando se redacta una pregunta debe pensarse en cómo se va a utilizar, en los análisis que se pretenden llevar a cabo.

Es posible que alguien se esté preguntando si no hubiera sido más sencillo poner una lista de asignaturas y pedirle al entrevistado que *marque* las que le obligan a leer. Varios argumentos justifican nuestra forma de actual:

- Tener que elegir una determinada asignatura en una lista de 43 complica la tarea del entrevistado en la medida que es más cómodo apuntar una –o varias– asignaturas que elegir de una lista tan numerosa. Dicho de otro modo, poner una lista de 41 asignaturas y decir que señale las correspondientes hace más "dura" la respuesta del cuestionario.
- Al entrevistar a personas de carreras muy diversas es prácticamente imposible recoger todas las asignaturas posibles
- Diferencia entre pregunta de recuerdo espontáneo y sugerido ¿Es lo mismo preguntar sobre las asignaturas que obligan a leer?, que ¿cuál/es de las asignaturas que te muestro obligan a leer? ¿Es lo mismo preguntar por los anuncios que viste anoche en el bloque publicitario de las 22 horas?, ¿que preguntar si en ese momento había anuncios de jabón Ariel?
  Las preguntas que no "sugieren" la respuesta (que asignaturas obligan a leer, que anuncios viste anoche) son conocidas como preguntas de recuerdo espontáneo, y suelen lograr menos respuestas que aquellas en las que se muestran las posibles respuestas. Ante esta situación conviene contar también con un código que indique el *no recuerdo*, la no respuesta porque no recuerda lo que se le pregunta. En este caso se ha elegido el número 96.

Otro aspecto importante a considerar es conocer la razón por la que se ha codificado esta pregunta con las últimas variables del libro de códigos (de la v50 a la v58), en vez de emplear el número correlativo a la pregunta 6. Cuando se redactó el cuestionario no se sabía el número de asignaturas que proponían libros de lectura obligatoria, y se consideró que bastaba con dejar cuatro respuestas. Al analizar los primeros cuestionarios respondidos se observó que el número de respuestas era mayor, por lo que fue necesario ampliar el número de respuestas. La mejor forma para ampliar el número de respuestas –sin que esto implique modificar todo el orden correlativo del cuestionario– es colocar las respuestas de esta pregunta al final del archivo de datos[34].

Finalizaremos la explicación de la codificación de respuestas señalando que en el libro de códigos del apartado 3.9 todos los "nuevos" valores aparecen en negrilla[35] para

34. Se trata de una pregunta multirespuesta categórica, con lo que es preciso tener en cuenta las recomendaciones esgrimidas unos párrafos más atrás.
35. No olvidemos que se trata de *convenciones* decididas por el investigador tras la elaboración del cuestionario.

diferenciarlos del resto de valores del cuestionario. De la variable V59, donde se recoge el "lugar de respuesta" (V59), hablaremos en el apartado 8. Terminada la explicación, llega la hora de ponerla en práctica revisando cada uno de los cuestionarios respondidos y anotando (cuando sea preciso), los valores de codificación. Éste es el objetivo del tercer ejercicio propuesto en los *materiales complementarios*.

Insistimos en la importancia de repasar y codificar cuestionario porque quizás nosotros no metamos los datos (grabación). La mayor parte de las veces se envía a una empresa especializada, o esta tarea puede ser realizada por otro miembro del equipo. En cualquier caso, aún cuando lo hagamos nosotros, el que introduce la información no debe *leer* el cuestionario sino sólo los datos (para meter la información con más rapidez). ¿Qué pasaría si, en vez de 40 tuviéramos 2.500 cuestionarios? De hecho, en los cuestionarios que se adjuntan en los *materiales complementarios* (carpeta *capítulo 3*) puede apreciarse que el día de la *edición* de cuestionarios se *coló* un error: Una persona (cuestionario número 115) ha cometido un error en la pregunta 19 (v43) al proporcionar dos respuestas: dice vivir con su padres y con su abuela. Considerando esta respuesta puede vivir sus padres, con su abuela, y con sus padres y abuela. Al desconocer la situación debemos codificarlo como mal respondida.

Terminaremos este apartado presentado las recomendaciones de Francisco Alvira (en su libro sobre la perspectiva general de la encuesta) para la realización del libro de códigos. Tras señalar que éste "depende, entre otras cosas, de los requerimientos del soporte informático y del análisis, pero sobre todo de los objetivos de la encuesta", señala varias indicaciones que deben ser consideradas en la elaboración del libro de códigos (Alvira, 2004: 55-56):

- "El orden del libro debe seguir el orden de las preguntas del cuestionario.
- Cada sistema de categorías debe incluir:
  - El número de la pregunta con el enunciado correspondiente,
  - descripción y título de la variable,
  - numeración de los campos,
  - y el valor numérico del sistema de categorías.
- Conviene siempre especificar categorías que preserven el máximo de la información original.
- Codificar la falta de información, sea por no respuesta o por no aplicable.
- Asegurarse que las categorías establecidas son exhaustivas y mutuamente excluyentes.
- Asignar códigos *convencionales* que se mantengan siempre: por ejemplo el 9, 99, 999... para la categoría "no contesta"; el 8, 88, 888... para la categoría "no sabe", etc. Estas *convenciones* implican también utilizar números correlativos cuando se trate de respuestas numéricas".

Recomendamos, en este punto de la exposición, comprobar hasta que punto el libro de códigos del apartado 9 cumple cada una de las recomendaciones especificadas por el profesor Alvira.

7. Introducción de datos (grabación)

Llega el momento de a introducir los datos, y lo haremos colocándolos directamente sobre el Editor de datos, dentro de la vista de datos. Pese a que el SPSS dispone de un programa de introducción de datos, el hecho que éste no esté incluido en la versión para estudiantes, unido a una escasa utilización del mismo nos lleva a introducir los datos directamente al *Editor de datos*. No obstante, es preciso señalar que numerosos investigadores realizan el proceso de introducción de datos utilizando programas de bases de datos. Este proceso de grabación de la información se realiza, normalmente, por duplicado y empleando personas distintas con el objetivo de minimizar los errores cometidos (Granero et al, 2001). Una vez finalizada la introducción de datos se comparan los archivos resultantes para analizar en detalle las diferencias.

Comenzaremos la introducción de datos por la primera celdilla de la izquierda, enmarcada con un borde más oscuro y cuyo nombre aparece en la parte superior izquierda del Editor, justo debajo de la barra de herramientas (figura 3.15). Introducido el valor correspondiente pulsaremos el tabulador o la *flecha derecha* de las teclas de dirección del teclado para validar este dato y dar paso a la siguiente variable (celdilla de la derecha). Es posible introducir los datos de izquierda a derecha, de derecha a izquierda, de arriba abajo, etc.; basta con pulsar la tecla correspondiente (flecha superior, inferior, derecha, izquierda) o situarnos con el ratón en la celdilla siguiente.

Al finalizar la introducción de todos los cuestionarios la orden *Archivo⇒Guardar* o "Ctrl+G" guardará el trabajo realizado; si bien la primera vez que se pulse solicitará un nombre para identificar el archivo. En este ejemplo el archivo recibirá el nombre "Práctica grabación" y deberá ser grabado en el dispositivo utilizado por cada lector (diskette, CD, salida USB, etc). Aunque no es estrictamente necesario guardar la información hasta que no se finalice la sesión (antes de cerrar el programa o apagar el ordenador), recomendamos guardarla cada vez que se introduce un nuevo cuestionario.

En el ejemplo que venimos desarrollando, al tomar la información desde un cuestionario respondido por un grupo de entrevistados, la introducción de datos se realizará por líneas, de izquierda a derecha. Conviene recordar que en el apartado 2.2, concretamente en la explicación del cuadro 2.2, se señaló que la información de cada

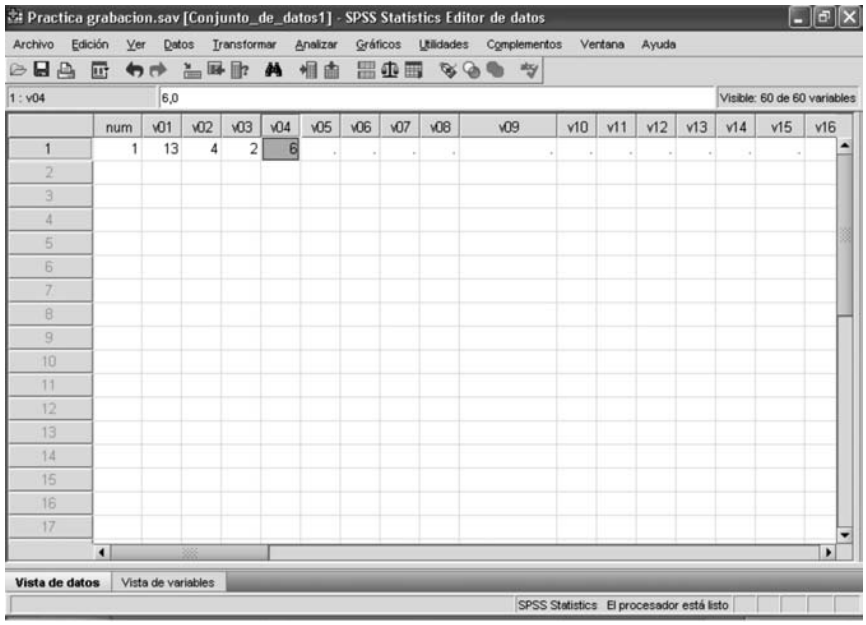


Figura 3.15. Introducción de datos.

entrevistado queda recogida en filas, mientras que en columnas se muestra cada una de las variables.

Utilizando el botón secundario del ratón en cualquier celdilla de la ventana de datos (pulsada la solapa *Vista de datos*) aparece el menú contextual mostrado en la figura 3.16, que facilita realizar determinadas acciones con los valores introducidos. Si en vez de hacerlo sobre la ventana de datos se realiza en la parte izquierda de la pantalla, en el área sombreada correspondiente al número de caso, se muestra un menú contextual que afecta a los casos introducidos, permitiendo *Cortar*, *Copiar*, Pegar, *Borrar* e *Insertar* un caso.

Tras la introducción y grabación de datos, llega el momento de realizar una revisión y depuración de la información recogida; tal y como se explicó en el primer capítulo. Antes de dar paso al proceso de depuración de datos creemos conveniente dedicar unas líneas a las posibilidades de edición y transformación de datos. En el segundo capítulo definimos la *edición de datos* como un conjunto de procesos diseñados y utilizados para detectar casos erróneos en datos de encuesta con el fin de corregirlos en la medida de lo posible. La edición y posterior transformación de datos en el SPSS se realiza principalmente utilizando los menús *Edición* y *Transformar*. Aunque el funcionamiento del programa SPSS se presenta con detalle en el siguiente capítulo, seña-

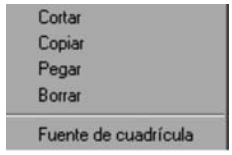


Figura 3.16. Menú contextual. Botón derecho del ratón sobre *Vista de datos*.

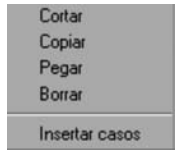


Figura 3.17. Menú contextual. Botón derecho del ratón sobre el número del caso, dentro de *Vista de datos*.

lar aquí que el menú *Edición* permite *Cortar*, *Copiar*, *Pegar* y *Borrar* los datos de las celdillas seleccionadas, así como la posibilidad de volver a la situación anterior si se ha cometido un error (*Deshacer*). La función *Buscar* permite localizar un determinado valor. Otros procedimientos de edición y transformación de datos, que forman parte del menú *Transformar*, serán explicados pormenorizadamente en el octavo capítulo.

8. Unión de archivos que contienen una estructura de información similar

A estas alturas del capítulo todos los lectores –tras codificar los cuestionarios respondidos– han creado un archivo de datos y han introducido las respuestas de los cuestionarios respondidos. Cuando explicamos la lógica de funcionamiento del texto (capítulo I) insistimos en la importancia de comparar los hábitos de lectura y dominio de equipos informáticos de esta promoción con los hábitos manifestados por promociones anteriores.

Para ello es necesario utilizar la operación *Fundir archivos*, que forma parte del menú *Datos*. Al poner el ratón sobre esta operación aparecen las dos opciones de unión de archivos considerando el elemento que tienen en común: casos o variables (figura 3.18). En el tema que nos ocupa unimos el cuestionario de esta promoción con el mismo cuestionario respondido por las promociones anteriores, de modo que se realiza la unión de casos.

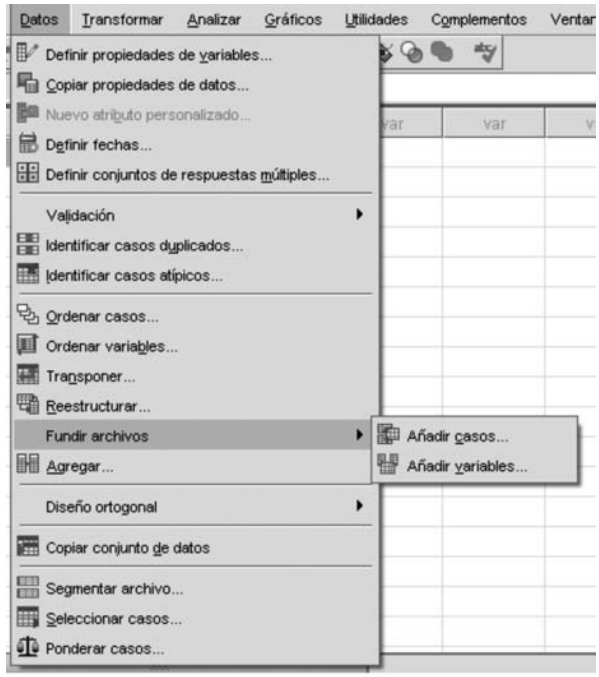


Figura 3.18. Opciones de Fundir Archivos: Añadir casos… y Añadir variables…

El proceso comienza abriendo el archivo de datos donde *copiar* los nuevos datos, que se denomina "Encuestas estudiantes (SEIS promociones).sav". Posteriormente se marca con el ratón el menú Datos⇒Fundir archivos⇒Añadir casos (figura 3.18) y aparece un cuadro de diálogo donde solicita el nombre del archivo a unir (figura 3.19). A continuación se selecciona el archivo recién creado ("Práctica grabación.sav") y, tras pulsar el botón *Continuar*, aparece el cuadro de diálogo de la figura 3.20 que informa de las variables desemparejadas y las variables que formarán el nuevo archivo de datos. En la esquina inferior izquierda se indica a que archivo pertenece la variable desemparejada, en este caso al conjunto de datos activo, esto es "Encuestas estudiantes (SEIS promociones).sav".

Para unir dos archivos es necesario que compartan el mismo número de variables, puesto que de lo contrario las variables *no compartidas* (desemparejadas) no formarán parte del archivo de datos resultante de la fusión. En este ejemplo existe una variable, v59, que está en el archivo de datos de otras promociones pero no en el archivo recién creado. ¿De qué se trata? Antes de unir dos archivos de datos es necesario analizar pormenorizadamente ambos archivos, comparando los libros de códigos o utilizando el menú *Utilidades⇒Información del Archivo*.

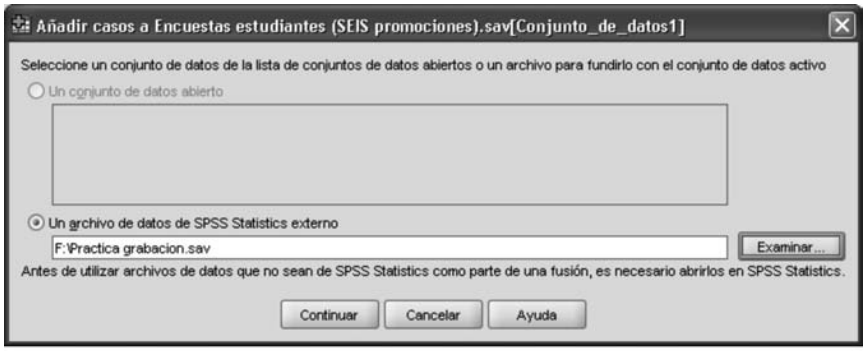


Figura 3.19. Fundir archivos, selección del archivo a unir.

Al abrir con SPSS el archivo "Encuestas estudiantes (SEIS promociones).sav" vemos un archivo exactamente igual que al elaborado al seguir las recomendaciones del capítulo, salvo dos pequeños detalles. El primero está referido al número de encuesta (variable num), ordenadas correlativamente del valor 1001 hasta el 1088, y a partir de este número *salta* hasta el 1101. Analizando el resto del archivo puede comprobarse que del 1159 vuelve a *saltar* hasta el 2001. Posteriormente del 2024 pasa hasta el 2101. Estas diferencias están desvelando –considerando el primer dígito– la promoción analizada, y –considerando el segundo dígito– de entrevistado (estudiantes de Sociología y sus amigos). Ahora bien, ¿qué información se ha introducido en esta variable en el archivo "Práctica grabación.sav"? El número de encuesta, que comenzaba por el cero en el caso de los cuestionarios realizado en clase, y por 1... en el caso de los cuestionarios respondidos por los amigos.

El segundo detalle detectado al observar ambos archivos es la presencia de una nueva variable en "Encuestas estudiantes (SEIS promociones).sav". Se trata de la variable v59, que recoge información sobre el "tipo de entrevistado" (estudiante de sociología ó amigo) considerando el lugar de respuesta: en los cuestionarios respondidos en clase aparecen las respuestas de los estudiantes de Sociología, mientras que los respondidos fuera de clase proporcionan información de los amigos.

Así, para realizar una correcta unión de archivos será necesario introducir en el archivo "Práctica grabación.sav" una nueva variable, denominada v59, que recoja información sobre dónde se ha respondido el cuestionario. Lógicamente, tras definir esta variable será necesario introducir los valores pertinentes[36]. Posteriormente se

36. Al tratarse de un archivo con pocos casos, la introducción de datos de esta nueva variable se realizará de forma manual caso a caso. En el capítulo VIII se explica como hacerlo de forma automática.

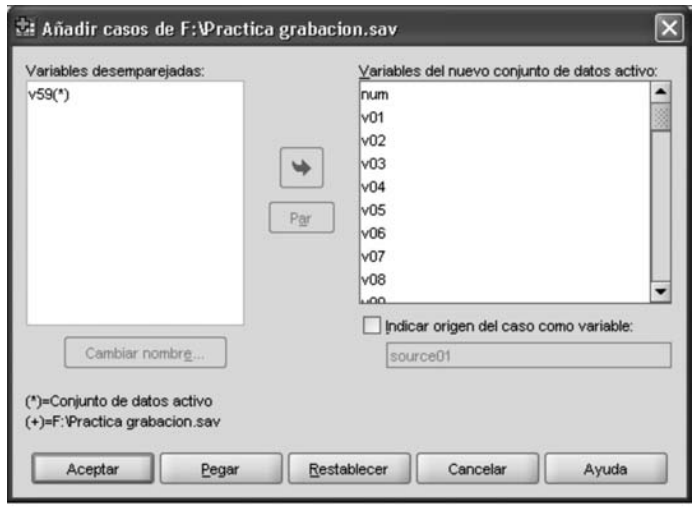


Figura 3.20. Añadir casos desde: variables desemparejadas, variables en el nuevo archivo.

procede con la fusión de archivos, logrando así un archivo con los hábitos de lectura y dominio de equipos informáticos de los estudiantes de Sociología de todas las promociones anteriores. Para terminar, guardaremos el archivo (con todos los casos) con el nombre "Encuestas estudiantes (SIETE promociones).sav".

Es importante que el investigador guarde su propia copia de este archivo, puesto que a partir de este momento todas las explicaciones deberán se aplicadas por los lectores a su archivo de datos, debiendo guardar todas las modificaciones introducidas a lo largo del texto.

Buscando fijar los conocimientos aprendidos en este apartado recomendados realizar los ejercicios propuestos en la carpeta capítulo 3 de los *materiales complementarios*, que proponen una aplicación práctica de los contenidos expuestos a lo largo del capítulo. Allí se presentan también diversos materiales para hacer prácticas (cuestionarios respondidos para codificar e introducir los datos, libro de códigos, archivos de datos, etc.).

9. Anexo 1: libro de códigos (inicial) del cuestionario "Encuestas estudiantes"

Pregunta	Variable Num	Código	Etiquetas Número de encuesta
P1	v01		Actividad, *fuera de casa*, que más te gusta hacer cuando dispones de tiempo libre[37]
		1	Beber, ir de copas
		2	Bailar
		3	Hacer deporte
		4	Ir de excursión
		5	Viajar
		6	Ir al cine
		7	Ir al teatro
		8	Ir a museos
		9	Ir a conciertos
		10	Leer libros
		11	Leer periódicos
		12	Leer revistas
		13	Practicar alguna afición o hobby
		14	Otras:
		15	Ninguna en particular
		16	**Quedar/salir con amigos/as**
		17	**Quedar/salir con el novio/a**
		18	**Ir al monte**
		98	**Mal respondida (más de una respuesta)**
		99	**No responde**
P2		v02	Actividad, *dentro de casa*, que más te gusta hacer cuando dispones de tiempo libre

37. Algunos investigadores no incluyen en el libro de códigos los valores de las preguntas parcialmente categorizadas, colocando en su lugar "precodificada". En esta pregunta, por ejemplo, se pondría "parcialmente recodificada" después del texto de la pregunta, mostrando únicamente los códigos que no están incluidos en el cuestionario: el 16, 17, 18, 98 y 99. El Centro de Investigaciones Sociológicas, por ejemplo, opera de esta forma. Personalmente creo que es mejor ponerlas porque de este forma el investigador puede prescindir del cuestionario.

Pregunta	Variable Num	Código	Etiquetas Número de encuesta
		1	Beber
		2	Bailar
		3	Ver la televisión
		4	Manejar el ordenador
		5	Jugar con videojuegos, playstation, etc.
		6	Dormir, descansar, no hacer nada
		7	Trabajar en las tareas del hogar
		8	Estudiar
		9	Escuchar música
		10	Leer libros
		11	Leer periódicos
		12	Leer revistas
		13	Practicar alguna afición o hobby
		14	Otras
		15	Ninguna en particular
		16	**Escribir**
		17	**Estar con mi familia**
		98	**Mal respondida (más de una respuesta)**
		99	**No responde**
P3	v03,v04		Situaciones que mejor definen tu actividad durante el tiempo libre
		1	Pasarlo bien sin hacer nada
		2	Hacer muchas cosas, estar activo, ir de un lado a otro
		3	Dedicarme a las personas más queridas
		4	Hacer cosas de mi trabajo que tengo pendientes
		5	Descansar, recuperar fuerzas
		6	Estar con la gente, charlar, tratar a los amigos
		7	Aburrirme
		8	Pensar, meditar
		9	Dedicarme tranquilamente a mis cosas, mis aficiones, deportes
		10	Otras

Pregunta	Variable Num	Código	Etiquetas Número de encuesta
		11	**Hacer cosas pendientes**
		98	**Mal respondida (más de dos respuestas)**
		99	**No responde**
P4	v05		Número de horas libres que dispones a la semana para tu ocio o diversión
		—	Horas anotadas
		0	Ninguna
		998	**Mal respondida**
		999	**No responde**
P5	v06		Número de libros, relacionados con tus estudios, leídos en el último año
		—	Libros anotados
		0	Ninguno
		98	**Mal respondida**
		99	**No responde**
P6	v07		Número de libros, relacionados con tus estudios, de lectura obligatoria
		—	Libros anotados
		0	Ninguno
		98	**Mal respondida**
		99	**No responde**
P8	v08		Número de libros leídos en este curso académico, exceptuando los relacionados con los estudios.
		—	Libros anotados
		0	Ninguno
		98	**Mal respondida**
		99	**No responde**

Pregunta	Variable Num	Código	Etiquetas Número de encuesta
P9	v09		Número de libros en el hogar
		—	Libros anotados
		0	Ninguno
		9.998	**Mal respondida**
		9.999	**No responde**
P10	v10		Número de periódicos de información general leídos semanalmente
		—	Periódicos anotados
		0	Ninguno
		98	**Mal respondida**
		99	**No responde**
P11	v11		Número de periódicos deportivos leídos semanalmente
		—	Periódicos anotados
		0	Ninguno
		98	**Mal respondida**
		99	**No responde**
P12	v12		Número de revistas leídas desde navidades
		—	Revistas anotadas
		0	Ninguna
		98	**Mal respondida**
		99	**No responde**
P13	v13		Tipo de revista (la última revista leída)
		01	Información general
		02	Corazón
		03	Moda
		04	Deportiva
		05	Economía
		06	Información televisión
		07	Profesional
		08	Decoración
		09	Erótica
		10	Motor

Pregunta	Variable Num	Código	Etiquetas Número de encuesta
		11	Pasatiempos
		12	Científicas (Muy Interesante, etc.)
		13	Viajes
		14	Ordenadores/informática
		15	Masculinas (Man, Men's Health, etc.)
		16	Femeninas (cosmopolitan, etc.)
		17	Musical
		18	Humorística
		19	Otras
		20	Ninguna
		21	Todas
		22	**Sindical**
		23	**Consumo**
		24	**Política**
		98	**Mal respondida (más de una respuesta)**
		99	**No responde**
P14	v14		Tiempo diario (en minutos) viendo la televisión en *días laborables*
		—	Minutos anotados
		0	Nada
		998	**Mal respondida**
		999	**No responde**
P15	v15		Tiempo (en minutos) viendo la televisión durante el fin de semana
		—	Minutos anotados
		0	Nada
		998	**Mal respondida**
		999	**No responde**
P16	v16		Presencia de vídeo o DVD en el hogar
		1	Si
		2	No
		98	**Mal respondida**
		99	**No responde**

Pregunta	Variable Num	Código	Etiquetas Número de encuesta
P16a	v17		Número de películas vistas en el último mes
		0	Ninguna
		97	**Si. Ha alquilado, pero no dice cuantas.**
		90	**No procede (filtrado)**
		98	**Mal respondida**
		99	**No responde**
P16b	v18		Día de la semana que vio por última vez una película de vídeo o DVD
		1	Sábado
		2	Domingo
		3	Otro día festivo
		4	Otro día no festivo
		90	**No procede (filtrado)**
		98	**Mal respondida**
		99	**No responde**
P16c	v19		Forma de visionado: sólo ó acompañado
		1	Solo
		2	Acompañado
		3	No recuerda
		90	**No procede (filtrado)**
		98	**Mal respondida**
		99	**No responde**
P16d	v20		Utilización del vídeo para “grabar” de la televisión.
		1	Si
		2	No
		90	**No procede (filtrado)**
		98	**Mal respondida**
		99	**No responde**
P17	v21		Presencia de ordenador en el hogar
		1	Si
		2	No

Pregunta	Variable Num	Código	Etiquetas Número de encuesta
		98	**Mal respondida**
		99	**No responde**
P17a	v22		Periféricos o dispositivos en el ordenador: IMPRESORA
		1	Si
		0	No/no responde
		90	**No procede (filtrado)**
		98	**Mal respondida**
P17a	v23		Periféricos o dispositivos en el ordenador: MODEM
		1	Si
		0	No/no responde
		90	**No procede (filtrado)**
		98	**Mal respondida**
P17a	v24		Periféricos o dispositivos en el ordenador: ALTAVOCES
		1	Si
		0	No/no responde
		90	**No procede (filtrado)**
		98	**Mal respondida**
P17a	v25		Periféricos o dispositivos en el ordenador: CÁMARA VIDEO
		1	Si
		0	No/no responde
		90	**No procede (filtrado)**
		98	**Mal respondida**
P17a	v26		Periféricos o dispositivos en el ordenador: LECTORA CD
		1	Si
		0	No/no responde
		90	**No procede (filtrado)**
		98	**Mal respondida**

Pregunta	Variable Num	Código	Etiquetas Número de encuesta
P17a	v27		Periféricos o dispositivos en el ordenador: GRABADORA CD
		1	Si
		0	No/no responde
		90	**No procede (filtrado)**
		98	**Mal respondida**
P17a	v28		Periféricos o dispositivos en el ordenador: LECTORA DVD
		1	Si
		0	No/no responde
		90	**No procede (filtrado)**
		98	**Mal respondida**
P17a	v29		Periféricos o dispositivos en el ordenador: GRABADORA DVD
		1	Si
		0	No/no responde
		90	**No procede (filtrado)**
		98	**Mal respondida**
P17b	v30		Acceso a Internet desde el hogar
		1	Si
		0	No
		90	**No procede (filtrado)**
		98	**Mal respondida**
P17c	v31-v36		Software y CD que tiene el ordenador del hogar
		1	Procesador de texto
		2	Hoja de cálculo
		3	Bases de datos
		4	Juegos y diversiones
		5	Enciclopedias y diccionarios
		6	Programas educativos y juegos para aprender

Pregunta	Variable Num	Código	Etiquetas Número de encuesta
		7	Programas prácticos para gestionar asuntos
		8	CD's culturales, música clásica, etc.
		9	Otros
		10	Ninguno
		11	No sabe
		12	**Más de seis**
		90	**No procede (filtrado y más de seis)**
		98	**Mal respondida**
		99	**No responde**
P18	v37		Frecuencia de utilización: ORDENADOR
		1	Todos o casi todos los días
		2	Dos o tres veces a la semana
		3	Una vez a la semana
		4	Menos de una vez a la semana
		5	Nunca o casi nunca
		98	**Mal respondida**
		99	**No responde**
P18	v38		Frecuencia utilización: CONEXIÓN INTERNET
		1	Todos o casi todos los días
		2	Dos o tres veces a la semana
		3	Una vez a la semana
		4	Menos de una vez a la semana
		5	Nunca o casi nunca
		98	**Mal respondida**
		99	**No responde**
P18	v39		Frecuencia utilización: CORREO ELECTRÓNICO
		1	Todos o casi todos los días
		2	Dos o tres veces a la semana
		3	Una vez a la semana
		4	Menos de una vez a la semana

Pregunta	Variable Num	Código	Etiquetas Número de encuesta
		5	Nunca o casi nunca
		98	**Mal respondida**
		99	**No responde**
P18	v40		Frecuencia de utilización: PROCESADOR TEXTO
		1	Todos o casi todos los días
		2	Dos o tres veces a la semana
		3	Una vez a la semana
		4	Menos de una vez a la semana
		5	Nunca o casi nunca
		98	**Mal respondida**
		99	**No responde**
P18	v41		Frecuencia de utilización: HOJA DE CÁLCULO
		1	Todos o casi todos los días
		2	Dos o tres veces a la semana
		3	Una vez a la semana
		4	Menos de una vez a la semana
		5	Nunca o casi nunca
		98	**Mal respondida**
		99	**No responde**
P18	v42		Frecuencia de utilización: BASE DE DATOS
		1	Todos o casi todos los días
		2	Dos o tres veces a la semana
		3	Una vez a la semana
		4	Menos de una vez a la semana
		5	Nunca o casi nunca
		98	**Mal respondida**
		99	**No responde**
P19	v43		Situación de residencia. Vive con:
		1	Con padres
		2	Con amigos en piso compartido

Pregunta	Variable Num	Código	Etiquetas Número de encuesta
		3	Con pareja
		4	**Otras situaciones: padre ó madre.**
		5	**Otras situaciones: familia propia**
		6	**Otras situaciones: hermanos**
		7	**Otras situaciones: madre, madre y hermanos**
		8	**Otras situaciones: tíos, tios y primos**
		9	**Otras situaciones: madre y abuela**
		10	**Otras situaciones: residencia estudiantes/colegio mayor**
		11	**Otras situaciones: solo**
		12	**Otras situaciones: con otros familiares**
		98	**Mal respondida**
		99	**No responde**
P20	v44		Titulación
		1	**LADE**
		2	**Licenciatura en Economía**
		3	**Diplomatura Estudios Empresariales**
		4	**Licenciatura en Sociología**
		5	**Diplomatura Relaciones Laborales**
		6	**Diplomatura Profesorado de EGB**
		7	**Diplomatura Trabajo Social**
		8	**Diplomatura en Enfermería**
		9	**Ingeniería Industrial**
		10	**Ingeniería Técnica Industrial**
		11	**Ingeniería Agrícola**
		12	**Ingeniería Técnica Agrícola**
		13	**Ingeniería Telecomunicaciones**
		14	**Licenciatura en Derecho**
		15	**Ingeniero Técnico Mecánico**
		16	**Licenciatura Comunicación Audiovisual**
		17	**Educación social**
		18	**Ingenier. Técnica Informática de Gestión**

Pregunta	Variable Num	Código	Etiquetas Número de encuesta
		19	**Licenciatura en Medicina**
		20	**Diseño Gráfico**
		21	**Ingeniería Técnica Electrónica**
		22	**LADE y Derecho**
		23	**Magisterio**
		24	**Biología**
		25	**Arquitectura**
		26	**Publicidad**
		27	**Arte dramático**
		28	**Otras**
		97	**Ninguna (no estudia)**
		98	**Mal respondida**
		99	**No responde**
P21	v45		Centro de estudios
		1	UPNA
		2	Universidad de Navarra
		3	UNED
		4	Otros
		98	**Mal respondida**
		99	**No responde**
P22	v46		Curso actual
		—	Curso anotado
		97	**Otros: Proyecto fin de Carrera, Postgrados, Oposiciones, etc.**
		98	**Mal respondida**
		99	**No responde**
P23	v47		Número de hermanos
		—	Número anotado
		0	No tiene hermanos (hijo único)
		98	**Mal respondida**
		99	**No responde**

Pregunta	Variable Num	Código	Etiquetas Número de encuesta
P23a	v48		Lugar ocupado dentro de los hermanos
		1	El mayor
		2	El segundo mayor
		3	Edad intermedia entre todos ellos
		4	El segundo más joven
		5	El más joven
		6	**Misma edad**
		90	**No procede (filtrado)**
		98	**Mal respondida**
		99	**No responde**
P24	v49		Género
		1	Hombre
		2	Mujer
		98	**Mal respondida**
		99	**No responde**
P7	v50-v58		Asignaturas que proponen libros de lectura obligatoria
		0	**Ninguna**
		1	**Sociología General**
		2	**Ciencia Política**
		3	**Economía política**
		4	**Historia Política y Social contemporánea**
		5	**Estadística aplicada a las Ciencias Sociales**
		6	**Estructura Social y Estr. Social de España**
		7	**Sociología de las Organizaciones**
		8	**Teoría Sociológica Clásica I**
		9	**Teoría Sociológica I**
		10	**Sistema Político Español**
		11	**Técnicas de Investigación Social**
		12	**Métodos y Técnicas Investigación Social I**

Pregunta	Variable Num	Código	Etiquetas Número de encuesta
		13	Historia de las Ideas Políticas
		14	Historia de Navarra
		15	Sociología de las Identidades Colectivas
		16	Sociología Comunicación y Opini. Pública
		17	Sociología Política
		18	Estructura Económica
		19	Economía de Navarra
		20	Génesis del Estado Moderno
		21	Creencias, Mitos y Supervivencias en Mundo Actual
		22	Teoría Sociológica Clásica II
		23	Teorías Sociológicas Actuales
		24	Antropología Social
		25	Técnic. Avanzadas de Investigación Social
		26	Teoría de la Población
		27	Teoría Sociológica II
		28	Filosofía y Metodología de las CC.SS.
		29	Sociología del Trabajo y del Ocio
		30	Organizaciones Sindicales y Patronales
		31	Política Social
		32	Sociología Rural
		33	Sociología Urbana
		34	Geografía Humana
		35	Ideologías Políticas Contemporáneas
		36	Procesos de Cambio Político en la Sociedad Contemporánea
		37	Sistema Político Español
		38	Métodos y Técnicas de Investig. Social II
		39	Métodos y Técnicas de Investig. Social III
		40	Psicología Social
		41	Introducción al Cambio Social

Pregunta	Variable Num	Código	Etiquetas Número de encuesta
		50	Otras 1: no pertenecientes al Plan de Estudios de la Licenciatura en Sociología: libre elección, otras carreras, etc.)
		51	Otras 2: No estudiantes de Sociología.
		90	No procede (se coloca cuando el entrevistado declara que ninguna asignatura le obliga a leer)
		96	No recuerdo
		97	No queda claro. No reproduce bien el nombre de la asignatura
		98	Mal respondida
		99	No responde
	V59		Lugar de respuesta
		1	Clase (Licenciados en Sociología)
		2	Fuera de clase

Capítulo IV

Introducción al SPSS

1. Objetivos didácticos del capítulo

En este capítulo se realiza una somera explicación del programa estadístico *SPSS*, centrada fundamentalmente en las características del mismo y en describir los elementos del menú principal. Es un capítulo de gran importancia en la medida que se presenta la terminología que será utilizada a lo largo de todo el texto.

El capítulo comienza con un breve recorrido histórico del desarrollo de los programas estadísticos empleados en el análisis de grandes cantidades de datos, para proceder a continuación con los motivos que nos llevaron a centrar este libro en el *Statistical Package for Social Sciences* (*SPSS*). Posteriormente se presentan las normas y rutinas elementales de funcionamiento, para centrar la atención en la ventana del menú principal, el *editor de datos* de SPSS.

2. Contextualización histórica

La gran ventaja que tiene el uso de *paquetes estadísticos* es la facilidad y rapidez en la realización de los cálculos precisos para el análisis, al tiempo de eliminan una gran cantidad de procesos repetitivos de conteo y procesamiento de datos, procesos susceptibles de generar numerosos errores cuando son realizados por personas. Su desventaja, por otra parte, es la necesidad de dedicar una cantidad de tiempo al aprendizaje del paquete estadístico si se desea conocer todas sus posibilidades y aprovechar toda su potencia.

Aunque el gran desarrollo de la tecnología de los programas de análisis de datos se inicia en la década de 1940, en 1885 Herman Hollerith elaboró un sistema de introducción de datos para la oficina del Censo de los Estados Unidos que ha seguido utilizándose hasta los años 50 del siglo XX. Los rudimentarios instrumentos creados por Hollerith para la introducción de datos pueden ser definidos como los antecedentes directos de los modernos paquetes estadísticos, si bien este proceso de evolución requiere tener en cuenta varios estadios: a principios de los años 60 los ins-

trumentos electromecánicos dan paso a los primeros ordenadores electrónicos con escasa capacidad de velocidad y memoria. En estos años se produce el desarrollo del lenguaje *Fortran*, y esto hace posible que un gran número de investigadores sociales comiencen a elaborar sus propios programas de análisis de datos.

Los primeros programas estadísticos tan sólo realizaban una operación (medias, tablas de contingencia, correlaciones, etc.), aunque cada una de estas operaciones fueron rápidamente introducidas en paquetes complejos que posibilitaban la utilización de un elevado número de técnicas de análisis de datos. En la década de 1970 aparece el término *paquete estadístico* cuando los investigadores sociales y otros científicos afines comienzan a demandar programas amplios para analizar datos. Con este fin se ponen en marcha en la Universidad de Chicago (www.uchicago.edu) y en la Universidad de Michigan (www.umich.edu) varios equipos de científicos para elaborar programas informáticos que permitan realizar de forma integrada operaciones variadas como modificar datos, crear índices, analizar datos, etc.

De todos los paquetes estadísticos existentes los más conocidos son el SAS (www.sas.com/es), el STATA (www.stata.com), el STATGRAPHICS (www.statgraphics.net) y el SPSS (www.spss.com y www.spss.com/es/). Un buen paquete estadístico se define por la facilidad en su utilización, las facilidades y controles en la introducción de datos (*data entry*), la manipulación de los datos para seleccionar subgrupos y crear nuevas variables, el análisis de datos (univariable, bivariable y multivariable) y la calidad de los resultados obtenidos. Si bien los programas anteriormente expuestos cumplen sobradamente estos requerimientos, otros programas menos conocidos (*BARWIN, GLIM, Microstat, Minitab, OSIRIS, SCSS, SPAD, SPLUS, SSNAP II, Statgraphics, Systat, etc.*) logran satisfacer al investigador al cumplir la mayor parte de las características destacadas.

Tras la exposición y enumeración de algunos de los programas disponibles se ha decidido utilizar el paquete estadístico SPSS porque este programa, junto con el SAS, son los paquetes más utilizados en la investigación con encuestas. El hecho que la primera versión del programa fuera realizada a mediados de la década de 1960 ha contribuido –sin duda alguna– a este amplio conocimiento del mismo. La facilidad del uso de las últimas versiones para Windows, unido a la explicación de determinados procesos estadísticos en los menús de ayuda, ha contribuido enormemente a realizar esta elección. Otra de las razones es la existencia de una versión para estudiantes que posibilita, a un precio muy asequible, que éstos puedan adquirirlo y familiarizarse con él. Han sido estas últimas razones las que nos han llevado a elegir el SPSS en vez del SAS.

En cuanto a su estructura el SPSS está dividido en diversos módulos[38], cada uno con un determinado tipo de procesos estadísticos. El más sencillo es el módulo *Base*

38. Entre otros: *Advanced Statistcis, Amos*, *Answertree*, *Categorias*, *Data Entry*, *Modelos de Regresión*, *Pruebas Exactas*, *Tablas*, *Tendencias*, *Qi-analyst*, *Samplepower*, etc.

que incluye toda la estadística descriptiva univariable y bivariable, exploración, tablas de contingencia, comparación de medias, test de hipótesis, análisis de varianza, correlación y regresión lineal, cálculo de proximidades, pruebas no paramétricas, análisis factorial, análisis de clasificación (cluster), análisis discriminante, cálculo de proximidades, y representaciones gráficas.

3. Comenzando a trabajar con SPSS. Rutinas elementales de funcionamiento

La sesión de trabajo con el paquete estadístico SPSS para Windows comienza haciendo un clic del ratón en el recuadro *Inicio* situado en la esquina inferior izquierda de la pantalla, a continuación la opción *Programas* (o todos los programas), carpeta *SPSS inc*, posteriormente *Statistics 17.0* y, por último, *SPSS Statistics 17.0*. Otra posibilidad, más utilizada por los *usuarios* habituales de la aplicación, es seleccionar un archivo de datos de SPSS –con el explorador de Windows– y hacer doble clic sobre éste. Sea cual sea la versión del SPSS instalada el programa se abrirá con el fichero de datos seleccionado.

Seguidamente se obtiene la pantalla mostrada en la figura 4.1, que permite comenzar a trabajar con SPSS ejecutando el tutorial, introduciendo los datos para el análisis, ejecutando una consulta de SPSS creada anteriormente, creando una nueva consulta, o abriendo una fuente de datos[39]. Recomendamos no utilizar ninguna de estas opciones, pulsando para ello el botón *Cancelar* situado en la parte inferior de la pantalla, y así visualizar el *Editor de datos* de SPSS; similar a una *hoja de cálculo* (figura 4.2), tal y como indicamos en la sección 3 del capítulo III.

El *Editor de datos de SPSS* es la pantalla principal del programa (figura 4.2), el marco de trabajo desde donde realizaremos todas las operaciones de análisis, por lo que será necesario proceder con la explicación de cada una de sus partes y funciones. En la parte superior izquierda aparece el icono SPSS, símbolo que estará presente en todos los archivos de datos creados con este programa. A la derecha se encuentra el nombre del archivo de datos activo, ninguno en un archivo recién abierto (figura 3.1), pero no así en el editor de datos mostrado en la figura 4.2. Esta información, el símbolo de SPSS y el nombre del archivo de datos, aparecerá también en la parte inferior de la pantalla del ordenador, en la barra de tareas.

39. En esta pantalla aparecen los archivos abiertos anteriormente, considerando tanto archivos de datos como de resultados.

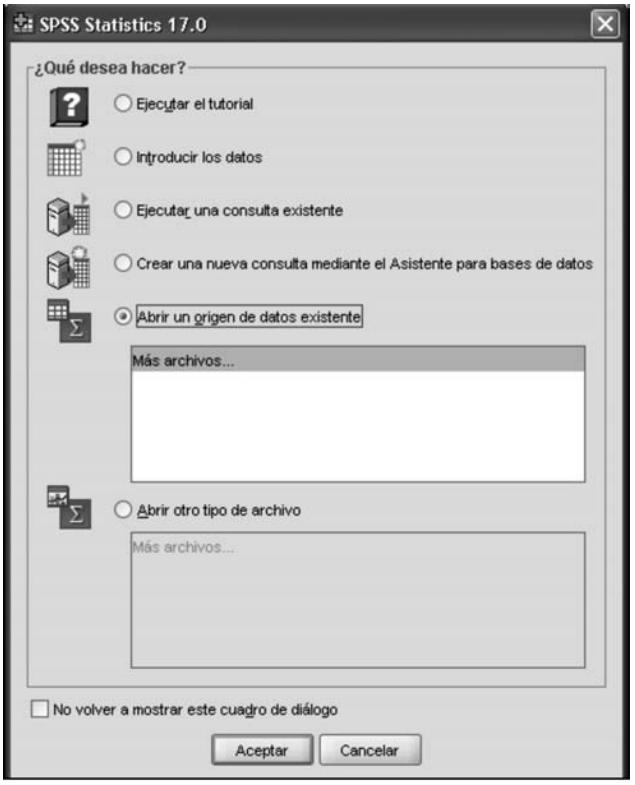
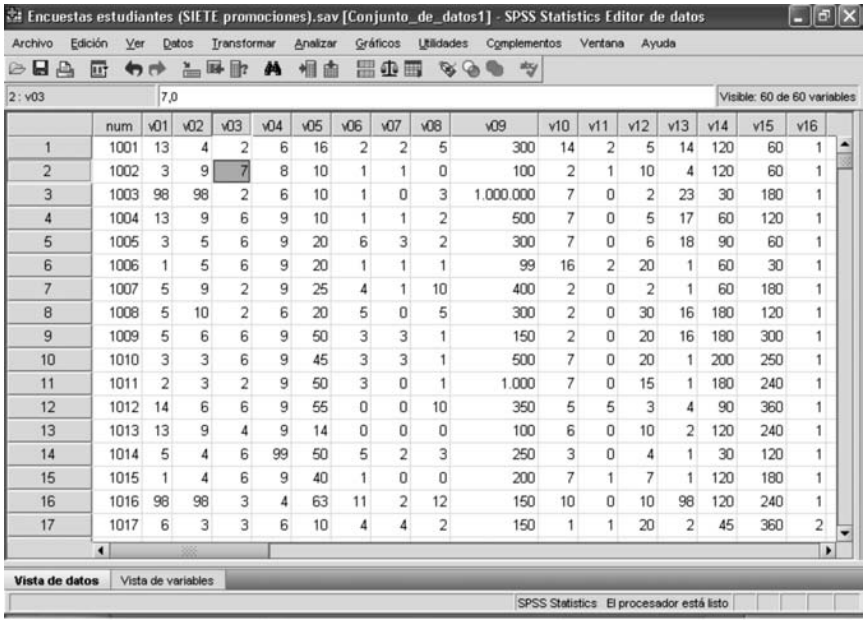


Figura 4.1. Pantalla de inicio de SPSS: menú de ayuda.

En la segunda línea del *Editor de datos* está situada la barra de menús, donde se encuentra el menú principal de SPSS con todas las funciones del programa agrupadas por temas: *Archivo*, *Edición*, *Ver*, *Datos*, *Transformar*, *Analizar*, *Gráficos*, *Utilidades*, *Complementos*, *Ventana* y *Ayuda*. Cada uno de éstos contiene distintos procedimientos para el análisis de la información incluida en el editor de datos, y será explicado con detalle en el próximo apartado.

La línea siguiente muestra la *barra de herramientas*, con iconos que permiten acceder rápidamente a los procedimientos más usuales de trabajo sin necesidad de acudir al menú y a los diferentes submenús.

Bajo la barra de herramientas se muestra el nombre de la celdilla donde se encuentra actualmente el cursor, y a la derecha (en el espacio en blanco) su valor. En la figura 4.2 puede apreciarse que el cursor está situado en la variable tres de la segunda línea (2:v03), y que esta celdilla tiene un valor de 7. Más abajo el nombre de las variables en el archivo de datos.



	num	v01	v02	v03	v04	v05	v06	v07	v08	v09	v10	v11	v12	v13	v14	v15	v16
1	1001	13	4	2	6	16	2	2	5	300	14	2	5	14	120	60	1
2	1002	3	9	7	8	10	1	1	0	100	2	1	10	4	120	60	1
3	1003	98	98	2	6	10	1	0	3	1.000.000	7	0	2	23	30	180	1
4	1004	13	9	6	9	10	1	1	2	500	7	0	5	17	60	120	1
5	1005	3	5	6	9	20	6	3	2	300	7	0	6	18	90	60	1
6	1006	1	5	6	9	20	1	1	1	99	16	2	20	1	60	30	1
7	1007	5	9	2	9	25	4	1	10	400	2	0	2	1	60	180	1
8	1008	5	10	2	6	20	5	0	5	300	2	0	30	16	180	120	1
9	1009	5	6	6	9	50	3	3	1	150	2	0	20	16	180	300	1
10	1010	3	3	6	9	45	3	3	1	500	7	0	20	1	200	250	1
11	1011	2	3	2	9	50	3	0	1	1.000	7	0	15	1	180	240	1
12	1012	14	6	6	9	55	0	0	10	350	5	5	3	4	90	360	1
13	1013	13	9	4	9	14	0	0	0	100	6	0	10	2	120	240	1
14	1014	5	4	6	99	50	5	2	3	250	3	0	4	1	30	120	1
15	1015	1	4	6	9	40	1	0	0	200	7	1	7	1	120	180	1
16	1016	98	98	3	4	63	11	2	12	150	10	0	10	98	120	240	1
17	1017	6	3	3	6	10	4	4	2	150	1	1	20	2	45	360	2

Figura 4.2. Menú principal: Editor de datos SPSS, vista de datos.

La mayor parte de la pantalla está ocupada por la *ventana de datos*, que recoge en filas la información de los casos y en columnas cada una de las variables medidas. En la parte inferior izquierda del editor de datos se presentan dos solapas que permiten acceder a la *Vista de datos*, con información sobre los datos introducidos; o *Vista de variables*, con información de las variables del fichero activo. No mostramos la vista de variables puesto que ha sido analizada con detalle en el tercer capítulo (sección 3.3).

A la derecha de estas solapas se encuentra la barra de desplazamiento izquierda/derecha que –al igual que el resto de aplicaciones que funcionan en el entorno Windows– permite un rápido desplazamiento para acceder a las últimas variables del archivo de datos (las colocadas más a la derecha). La flecha de desplazamiento, situada a continuación, permite un desplazamiento más pausado, variable a variable. En el margen derecho de la pantalla aparece la barra de desplazamiento vertical/horizontal, junto con la correspondiente flecha de desplazamiento.

En la última línea de la pantalla está situada la barra de estado, en la figura 4.2 con la leyenda *SPSS El procesador está listo*, y ofrece información sobre:

- Estado del comando: recuento de casos con el número de casos procesados en cada momento, y número de iteraciones en los procedimientos estadísticos que lo precisan.

- Estado del filtro: aparece el mensaje *Filtro activado* cuando se ha seleccionado una muestra aleatoria de casos, o se está trabajando con una parte de los datos (se analizará en detalle en el octavo capítulo, apartado 8).
- Estado de ponderación: el mensaje *Ponderación Activada* especifica que se está utilizando una variable para *ponderar* los casos (se muestra una aplicación práctica de este proceso en el sexto capítulo, apartado 7).
- Estado de división de archivos: el término *Segmentar por...* señala que el archivo de datos ha sido segmentado en diferentes grupos por la variable indicada (se presenta un ejemplo de utilización en el capítulo VIII, apartado 9).

4. Ventana del menú principal

Dedicaremos este apartado a explicar las distintas funciones disponibles en la barra de menús del programa, segunda línea del *Editor de datos*. Mediante una serie de menús despegables la barra del menú del programa permite acceder a las principales funciones del programa agrupadas temáticamente: Archivo, Edición, Ver, Datos, Transformar, Analizar, Gráficos, Utilidades, Complementos, Ventana y Ayuda. En la figura 4.3 se muestra el menú Archivo, que será utilizado para explicar las convenciones presentes en todos los menús:

- La letra subrayada en cada una de las funciones indica que ésta se activa pulsando la tecla correspondiente, siempre que el menú esté desplegado. Cuando el menú no está desplegado puede activarse esta función pulsando conjuntamente *Alt* y la tecla subrayada.
- El signo ▸ indica que hay más opciones, a las que se accede al colocar el cursor de ratón sobre la función.
- Los puntos suspensivos indican que marcando esta función se accede a un cuadro de diálogo.
- Una notación a la derecha de cada submenú, como sucede en *Guardar* o *Imprimir*, indica que se puede acceder a esa función sin necesidad de desplegar el menú. Desde cualquier parte del programa, pulsando Ctrl+P permitirá acceder al cuadro de diálogo *imprimir*.
- Las funciones con menor color, como *Cerrar*, *Guardar todos los datos*, *Detener el procesador* (en la figura 4.3), o *Deshacer*, *Rehacer*, *Pegar* y *Pegar variables* (en la figura 4.4 de la página 111), no están disponibles en este momento. Lógicamente, la función "Detener procesador" cambia de color en el momento que se procesa alguna orden estadística.

Figura 4.3. Editor de datos SPSS: menú Archivo.

- En determinadas funciones aparece a la izquierda el signo ✓, que indica que la función está activada. Esto implica, en la figura 4.4, que se encuentran activadas la *Barra de estado* y la *Cuadrícula*.

Explicadas las convenciones comunes a todos los menús despegables, es el momento de analizar las funciones presentes en cada uno. El menú ***Archivo*** se utiliza para todo lo relacionado con operaciones de archivos de datos, resultados, o sintaxis de SPSS (figura 4.3). Está formado por los siguientes submenús:

- Nuevo: crea un nuevo archivo de datos, de sintaxis, resultados o proceso.
- Abrir: abre archivo de datos, de sintaxis, resultados o proceso.
- Abrir base de datos: crea, edita y ejecuta consultas a bases de datos.
- Leer datos de texto: abre archivos de texto.
- Cerrar: cierra el programa.
- Guardar: guarda todos los archivos abiertos.
- Guardar como: guarda el archivo actual con otro nombre, o con distinto formato.

- Guardar todos los datos.
- Exportar a base de datos.
- Marcar archivo como de solo lectura.
- Cambiar nombre del conjunto de los datos.
- Mostrar información del archivo de datos: muestra información de las variables incluidas en los archivos de datos, tanto del archivo de trabajo como de un archivo externo. Información proporcionada: tipo de variable, etiqueta, valores definidos como perdidos, nivel de medida y etiquetas de valores. Esta información se presenta en el editor de resultados, en formato texto.
- Caché de los datos: crea memoria para los datos que se introduzcan.
- Detener procesador: interrumpe el procesamiento de SPSS.
- Cambiar servidor: cambia el servidor al que se está conectado.
- Presentación preliminar: muestra, en pantalla completa, la selección realizada para imprimir.
- Imprimir: imprime la ventana actual.
- Datos usados recientemente: muestra los archivos de datos usados recientemente.
- Archivos usados recientemente: muestra los archivos usados recientemente.
- Salir: sale de SPSS.

Con el menú ***Edición*** se accede a las operaciones de cortar, copiar, pegar y borrar partes de un archivo de datos (Figura 4.4):

- Deshacer: deshace la última acción realizada.
- Rehacer: rehace la última acción deshechada.
- Cortar: corta la selección y la almacena en el portapapeles.
- Copiar: copia la selección y la almacena en el portapapeles.
- Pegar: pega – en la ubicación del cursor– la selección almacenada en el portapapeles.
- Pegar variables: pega la variable del portapapeles, en la ubicación del cursor.
- Borrar: elimina la selección realizada.
- Insertar variable: inserta una variable en cualquier lugar de un fichero de datos, a la izquierda de la celdilla donde está situado el cursor.
- Insertar caso: inserta un nuevo caso en el editor de datos, en la fila anterior donde está situado el cursor.
- Buscar: busca los datos especificados.
- Reemplazar: reemplazo de valores.
- Ir a caso: permite situarse en un determinado caso. Si antes de pulsar “ir a caso” se selecciona una variable haciendo un clic de ratón en cualquier lugar de ésta, el cursor se desplazará al valor de este sujeto en la variable seleccionada.

Figura 4.4. Editor de datos: menú Edición.

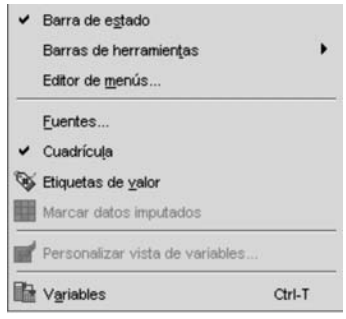


Figura 4.5. Editor de datos: menú Ver.

- Ir a variable: desplazamiento a una determinada variable.
- Opciones: opciones de tablas, gráficos, procesos, etc.

Dentro del menú ***Ver*** (figura 4.5) se muestran diversas funciones relacionadas con la presentación del Editor de datos:

- Barra de estado: activa (✔) y desactiva la barra de estado.
- Barras de herramientas: activa y desactiva la barra de herramientas.
- Editor de menús: muestra los menús disponibles.
- Fuentes: permite cambiar estilos y tamaños para las fuentes.
- Cuadrícula: activa y desactiva la cuadrícula del editor de datos.
- Etiquetas de valor: muestra las etiquetas de las variables en vez de los datos numéricos.
- Variables: activa la vista variables (o de datos) en el editor de datos. Se produce el mismo efecto que cuando se pulsa la solapa "vista de variables" o "vista de datos" situada en la parte inferior izquierda del editor de datos.

En el menú ***Datos*** (figura 4.6) se presentan diversas operaciones de definición y transformación de archivos de datos:

- Definir propiedades de variables: ayuda en el proceso de creación de etiquetas de valor para variables categóricas. Explora los datos y enumera los valores de datos únicos, identifica valores sin etiquetas y ofrece etiquetas automáticas, y permite copiar etiquetas de valor a otras variables.

- Copiar propiedades de datos: copia, en el archivo de datos actual, las propiedades de un conjunto de datos o variables de otro archivo. Es posible también copiar las propiedades de una variable a otra dentro del mismo archivo de datos.
- Definir fechas: permite la definición de valores fecha.
- Definir conjuntos de respuestas múltiples: define conjuntos de respuestas múltiples.
- Validación: propuesta de reglas de validación de datos.
- Identificar casos duplicados: definición de "dato duplicado", y localización en el archivo de datos.
- Identificar casos atípicos.
- Ordenar casos: ordena los casos en orden ascendente o descendente según los valores de la variable seleccionada. Opción disponible pulsando el botón secundario (derecho) del ratón sobre el nombre de la variable en la *vista de datos*. Supongamos que seleccionamos la variable "v09" del fichero mostrado en la figura 4.2 y se pide una clasificación según un orden ascendente: este proceso consistirá en poner en primer lugar al individuo cuyo valor en "v09" sea menor, por ejemplo 1, después 2, 3, y así sucesivamente (se explica con detalle en el capítulo VI, apartado 3).
- Ordenar variables: ordena las variables del Editor de datos según el criterio fijado.
- Transponer: cambia la definición del fichero de datos, considerando las variables como casos y los casos como variables. Pueden seleccionarse todas las variables o únicamente algunas de ellas. Las variables no seleccionadas no aparecerán en el nuevo archivo creado.
- Reestructurar: reestructura el archivo de datos, presentando tres opciones: reestructuración de variables seleccionadas en casos, casos seleccionados en variables, y transposición de todos los datos.
- Fundir archivos: permite la unión de archivos que contienen los mismos sujetos pero distintas variables (opción *Añadir Variables*), o que contienen las mismas variables pero distintos sujetos (*Añadir Casos*).
- Agregar: crea nuevos archivos con datos agregados.
- Diseño ortogonal: diseña y muestra diseños factoriales ortogonales.
- Copiar conjunto de datos: copia el archivo de datos activo a un nuevo archivo.
- Segmentar archivo: permite separar archivos según ciertos criterios.
- Seleccionar casos: selecciona unos determinados casos, aquellos que cumplen las condiciones especificadas.
- Ponderar casos: ponderación de casos.

Figura 4.6. Editor de datos: menú Datos.

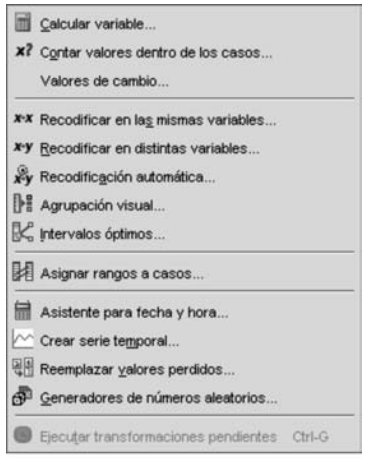


Figura 4.7. Editor de datos: menú Transformar.

El menú datos, utilizado para llevar a cabo determinadas operaciones con los datos, se complementa con ***Transformar*** (figura 4.7), donde se recogen diversos procedimientos de transformación de variables:

- Calcular variable: realización de cálculos
- Contar valores dentro de los casos: crea una variable que cuenta las apariciones de determinados valores en una lista de variables.
- Valores de cambio: crea nuevas variables que contienen los valores de variables existentes de casos anteriores o posteriores
- Recodificar en las mismas variables: unión de los valores de una variable en la misma variable.
- Recodificar en distintas variables: unión de los valores de una variable en otra variable.
- Recodificación automática: transforma los valores numéricos y de cadena en valores enteros consecutivos.
- Agrupación visual: transforma variables cuantitativas en cualitativas. Crea las nuevas variables utilizando percentiles, de modo que cada grupo contiene un número de casos similar.
- Intevalos óptimos: discretiza variables de escala.
- Asignar rangos a casos: crea nuevas variables formadas por rangos.

- Asistente para fecha y hora: asistente para el tratamiento de fechas y horas en SPSS.
- Crear serie temporal: crea nuevas variables basadas en funciones de variables de series temporales numéricas existentes.
- Reemplazar valores perdidos: reemplaza no respuestas y otros valores *missing*.
- Generador de números aleatorios: elaboración de números aleatorios, estableciendo un valor de secuencia de inicio de modo que pueda reproducir una secuencia de números aleatorios.
- Ejecutar transformaciones pendientes: lleva a cabo las transformaciones que están en espera.

Por su parte, dentro del menú ***Analizar*** se muestran todos los procedimientos estadísticos presentes en el programa (figura 4.8). Los procedimientos estadísticos disponibles dependen de los módulos adquiridos. Como ya hemos señalado más atrás este texto utilizará tan sólo una pequeña parte de los procedimientos estadísticos disponibles, si bien procederemos a la enumeración de todos:

- Informes: cubos OLAP, resúmenes de casos, e informes de estadísticos.
- Estadísticos descriptivos: Frecuencias, Descriptivos, Explorar, y Tablas de contingencia y Razón.
- Tablas: tablas de frecuencia y otros tipos de tablas.
- Análisis de RFM: técnica utilizada para identificar a clientes actuales que tienen más posibilidades de responder a una nueva oferta.
- Comparar medias: comparación de medias, prueba T de Student, análisis de la varianza.
- Modelo lineal general: análisis de varianza univariante, multivariante y de medidas repetidas.
- Modelos lineales generalizados: ampliación del modelo lineal general.
- Modelos (lineales) mixtos.
- Correlaciones: correlaciones parciales y bivariadas. Cálculo de distancias.
- Regresión: regresión lineal, no lineal, curvilínea, logística, ordinal, probit, etc.
- Loglineal: modelos logarítmicos lineales.
- Redes neuronales: análisis de redes neuronales *Perceptrón multicapa* y *Función de base radial*.
- Clasificar: análisis discriminante y análisis de conglomerados.
- Reducción de dimensiones: análisis de correspondencias, factorial y escalamiento óptimo.
- Escala: escalamiento multidimensional y análisis de la fiabilidad.
- Pruebas no paramétricas: Chi-cuadrado, binomial, rachas, etc.
- Predicciones: modelos ARIMA, autorregresión, suavizado y descomposición estacional.

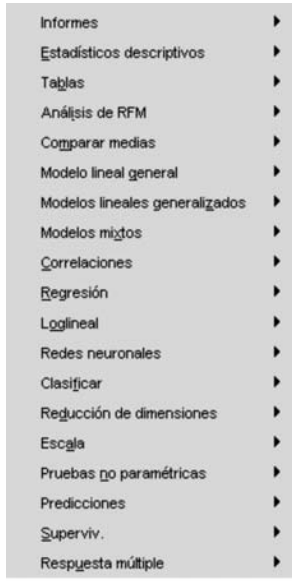


Figura 4.8. Editor de datos SPSS: menú Analizar.

- Supervivencia: tablas de mortalidad, Kaplan-Meier, y regresión de Cox.
- Respuestas múltiples: definir conjuntos de respuestas mútiples, frecuencias y tablas de contingencia.

El menú ***Gráficos*** se ocupa, lógicamente, de la presentación de resultados en dispositivos gráficos. En la figura 4.9 se muestran los gráficos disponibles. El programa dispone de un generador de gráficos, gráficos de Barras, Barras 3-D, Líneas, Áreas, Sectores, Máximos y mínimos, Diagramas de caja, Barras de error, Pirámide de población, gráficos de Dispersión/Puntos, Histograma, y gráficos interactivos.

Los siguientes menús reciben el nombre de *utilidades*, *complementos*, *ventana* y *ayuda*. ***Utilidades*** comprende diversas operaciones no clasificadas en otro sitio como informar sobre variables, sistemas de gestión de resultados, añadir comentarios al archivo de datos, definir y usar conjuntos de variables, ejecución de procesos, y edición de menús (figura 4.10). De todas éstas, las más interesante –para nuestros objetivos– es la primera. *Variables* proporciona información sobre cada variable, concretamente el tipo de variable, etiqueta, valores definidos como perdidos, nivel de medida y etiquetas de valores.

El menú ***Ventana*** se utiliza para minimizar SPSS, cambio y selección de ventanas abiertas. Esta versión del SPSS opera con siete tipos de ventanas: editor de datos

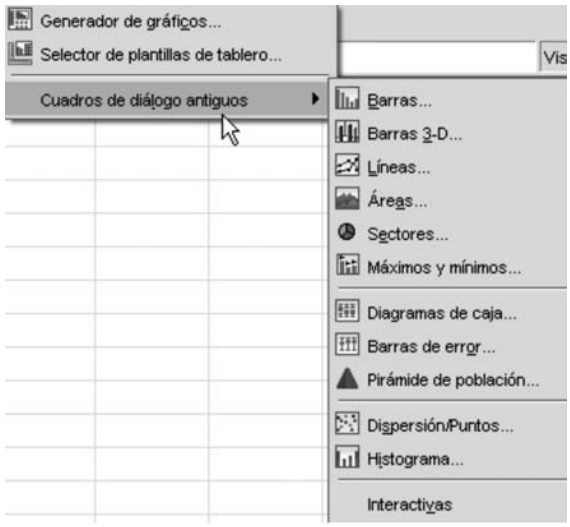


Figura 4.9. Editor de datos SPSS: menú Gráficos.

Figura 4.10. Editor de datos SPSS: menú Utilidades.

(varios archivos), visor de resultados, editor de tablas pivote, editor de gráficos, editor de resultados de texto, editor de sintaxis, y editor de procesos. La mayor parte de éstos serán explicados en los siguientes capítulos. Por último aparece el menú de *Ayuda*; que se presenta clasificada por Temas: dispone de un Tutorial, muestra la Guía de sintaxis de los comandos, presenta un Asesor estadístico y permite acceder –vía Internet–a la Página principal de SPSS.

Junto con las distintas funciones disponibles en la barra de menús del programa, segunda línea del Editor de datos, es importante explicar las funciones de la barra de herramientas, con iconos que permiten acceder rápidamente a los procedimientos más usuales de trabajo (esto es, sin necesidad de acudir al menú principal y a los diferentes submenús). La barra de herramientas está colocada en la tercera línea del Editor de datos de SPSS, como pudo verse en la figura 4.2, entre la barra del menú principal y el nombre de la celdilla donde se encuentra el cursor. Para utilizar cualquiera de las opciones que presenta la barra de herramientas basta hacer un clic sobre el recuadro correspondiente. El programa proporciona una breve descripción de la función de cada icono situando el puntero del ratón sobre él.

Figura 4.11. Barra de herramientas del Editor de datos de SPSS.

En la figura 4.11 se muestra la barra de herramientas del menú principal. Comenzando la explicación de izquierda a derecha, cada uno de sus iconos permiten acceder a las siguiente funciones:

Abrir documentos de datos.

Guardar documento. Si es la primera vez que se hace, mostrará el cuadro de diálogo para poner un nombre.

Imprimir documento activo en pantalla, siendo posible seleccionar una parte del documento.

Rellamada a los cuadros de diálogo, mostrando las últimas operaciones realizadas con el programa. Basta con hacer clic en una de ellas para que aparezca en pantalla el cuadro de diálogo correspondiente.

Deshacer: deshace la última acción realizada.

Rehacer: rehace la última acción deshechada

Ir a caso.

Ir a variable.

Información sobre la variable seleccionada

Buscar.

Insertar caso en el archivo de datos.

Insertar variable.

Segmentar archivo.

Ponderar casos

Seleccionar casos.

Muestra en el archivo de datos las etiquetas de valor en vez de los códigos numéricos.

Utilizar conjunto de variables ya predefinido para utilizarlo en el análisis.

Mostrar todas las variables.

Corregir ortografía.

Capítulo V

Importación de archivos creados por otros

1. Objetivos didácticos del capítulo

En el tercer capítulo se ha explicado como crear archivos de datos con información obtenida de la realidad social, en este caso considerando las respuestas a un cuestionario sobre prácticas de ocio, hábitos de lectura y dominio de equipos informáticos. Aunque esta es la situación más común, en numerosas ocasiones los investigadores –en vez de crear un archivo de datos– utilizan archivos existentes.

La mayor parte de las veces estos archivos han sido creados por otros investigadores utilizando los programas informáticos con más difusión. En este capítulo se explica como recuperar con el SPSS archivos elaborados con la hoja de cálculo *Microsoft Excel*, la base de datos *Microsoft Access*, y archivos texto en *ANSI* ó *ASCII*. Téngase en cuenta que en un proceso de investigación normalmente se opta por crear un archivo de datos (capítulo tres) o por utilizar un archivo existente (capítulo actual); pero que en raras ocasiones se utilizan ambas estrategias.

De modo que en las investigaciones basadas en archivos de datos *realizados por otros* no es necesario aplicar lo visto en el tercer capítulo, especialmente en los apartados 3.6, 3.7 y 3.8. Visto así, este capítulo *sustituiría*[40] a la *definición de variables y elaboración de un archivo de datos* (capítulo tres). En caso de que fuera así ambos capítulos debieran haberse colocado seguidas en el texto, si bien no se ha procedido de esta forma para que los lectores "descansen" tras el capítulo tres (formada por 51 páginas) con un sencillo capítulo cuarto (de 6 páginas) donde se explica el funcionamiento básico del programa estadístico SPSS. Otra razón que justifica situar este capítulo después de la *introducción al SPSS* es que la recuperación de archivos elaborados por otros precisa un mayor dominio del programa SPSS; esto es, conviene que el lector esté más familiarizado con la utilización del programa.

40. Más adelante veremos que –más que una *sustitución*– se trata de una complementación con el tercer capítulo.

2. Lectura–recuperación de archivos realizados con hojas de cálculo

La recuperación de archivos realizados con la hoja de cálculo Excel es tremendamente sencilla, puesto que SPSS reconoce directamente los archivos creados por este programa. Utilizando el menú *Abrir archivo* es posible conocer los archivos reconocidos directamente por el SPSS, accediendo con el ratón al menú despegable "Archivos de tipo". En la parte inferior de la figura 5.1 se muestran algunos de los archivos reconocidos por SPSS.

Una vez seleccionada la extensión "*.xls", que es la utilizada por archivos Excel, el programa mostrará en pantalla los archivos creados por esta aplicación: "EJEMPLO1-ESTUDIANTES... (EXCEL)", y "EJEMPLO-2 CONSUMO (EXCEL)"; tal y como se muestra en la figura 5.2. El primero es el archivo "Encuestas estudiantes 2002_03.sav", elaborado en el tercer capítulo y con el que se propone realizar un ejercicio en los *materiales complementarios* (web), carpeta denominada *capítulo 5*.

El archivo "EJEMPLO-2 CONSUMO (EXCEL)" son los resultados de un estudio sobre consumo respondido por 900 entrevistados y cuyo cuestionario se muestra en los *materiales complementarios*. En la figura 5.3 se muestra el archivo original, recuperado con el programa Microsoft Excel.

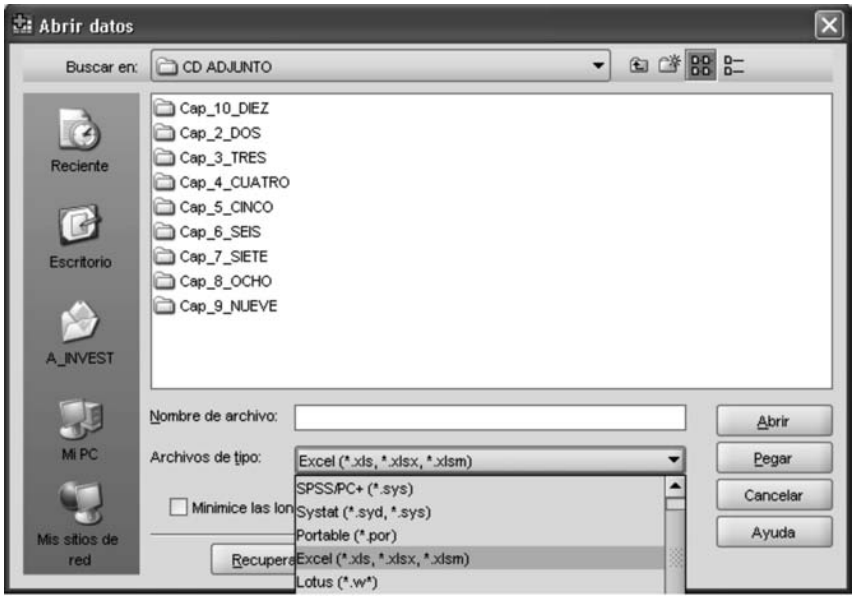


Figura 5.1. Abrir archivo datos: archivos reconocidos por SPSS.

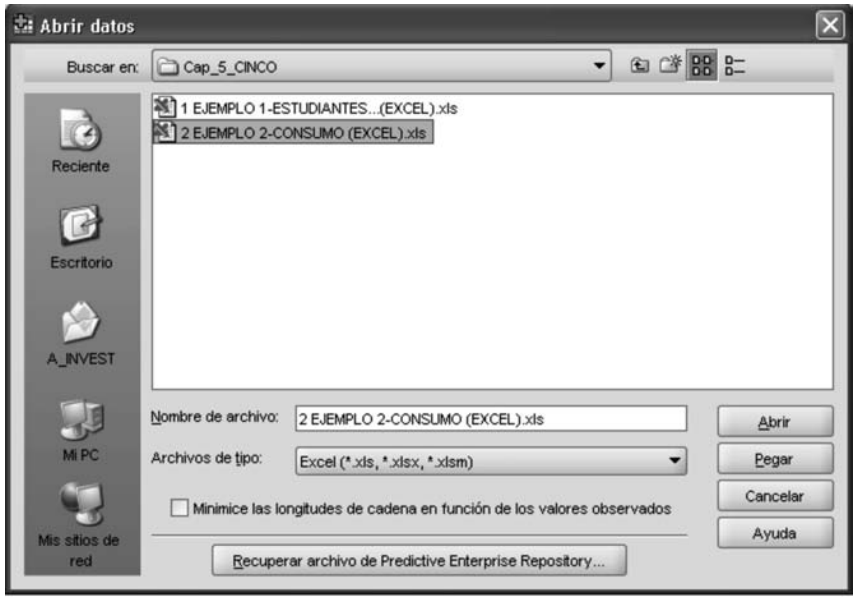
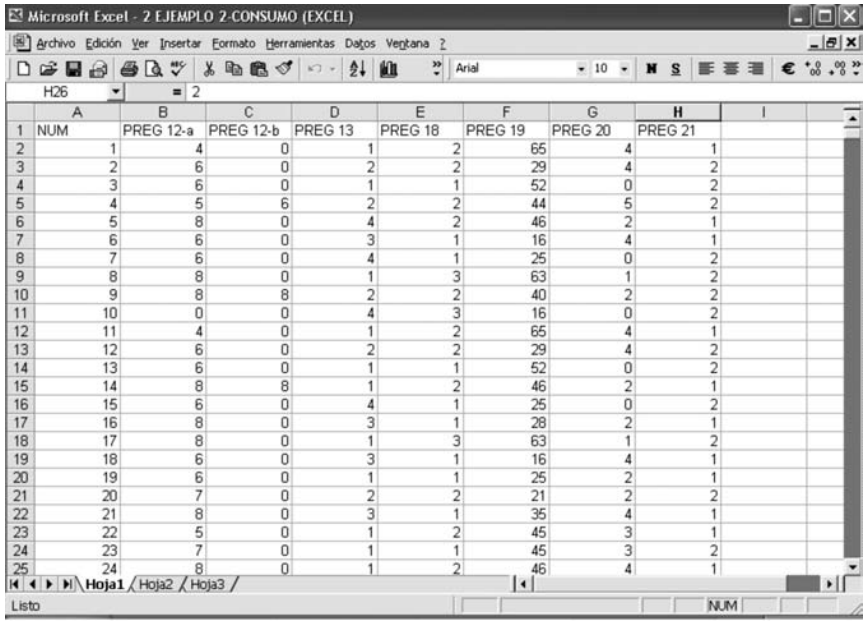


Figura 5.2. Abrir archivo Excel.



	A	B	C	D	E	F	G	H	I
1	NUM	PREG 12-a	PREG 12-b	PREG 13	PREG 18	PREG 19	PREG 20	PREG 21	
2	1	4	0	1	2	65	4	1	
3	2	6	0	2	2	29	4	2	
4	3	6	0	1	1	52	0	2	
5	4	5	6	2	2	44	5	2	
6	5	8	0	4	2	46	2	1	
7	6	6	0	3	1	16	4	1	
8	7	6	0	4	1	25	0	2	
9	8	8	0	1	3	63	1	2	
10	9	8	8	2	2	40	2	2	
11	10	0	0	4	3	16	0	2	
12	11	4	0	1	2	65	4	1	
13	12	6	0	2	2	29	4	2	
14	13	6	0	1	1	52	0	2	
15	14	8	8	1	2	46	2	1	
16	15	6	0	4	1	25	0	2	
17	16	8	0	3	1	28	2	1	
18	17	8	0	1	3	63	1	2	
19	18	6	0	3	1	16	4	1	
20	19	6	0	1	1	25	2	1	
21	20	7	0	2	2	21	2	2	
22	21	8	0	3	1	35	4	1	
23	22	5	0	1	2	45	3	1	
24	23	7	0	1	1	45	3	2	
25	24	8	0	1	2	46	4	1	

Figura 5.3. Archivo Excel (formato original).

Volviendo de nuevo al cuadro de diálogo de la figura 5.2, haciendo doble clic sobre este archivo emerge el cuadro de diálogo de la figura 5.4 donde el programa pregunta al usuario si desea recuperar el nombre de las variables de la primera fila de datos para que se utilicen como nombres de variables. Cuando los archivos de Excel disponen de varias hojas de datos éstas aparecen en la ventana *Rango*, y el usuario puede seleccionar la/s hoja/s de datos a recuperar. En este caso, al tener una hoja, esta ventana aparece vacía.

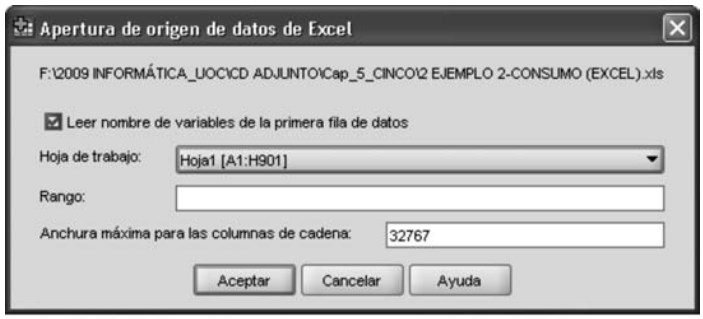
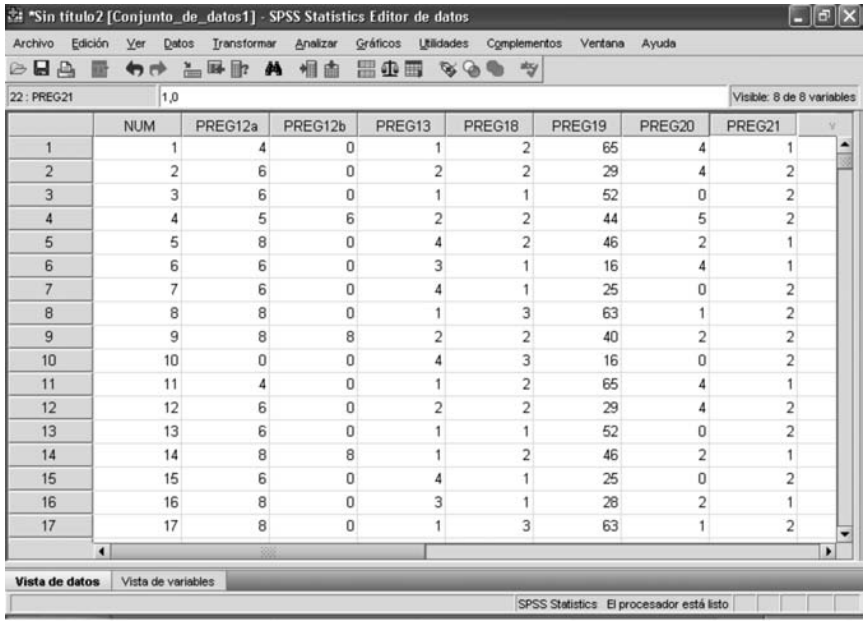


Figura 5.4. Abrir archivo Excel: opciones.

Tras pulsar *Aceptar* se lleva a cabo la recuperación de la información (ver figura 5.5). A continuación deberá procederse con la definición del formato de las variables (tipo, anchura, etiquetas, valores perdidos, columnas, alineación, etc.), tal y como se explicó en el apartado 3.3. De lo que se desprende que este capítulo no sustituye al tercero, sino que lo complementa.

A continuación deberán solicitarse las frecuencias –comparándolas con el de libro de códigos– para comprobar que el archivo se ha recuperado adecuadamente. Para realizar esta comprobación es necesario utilizar el procedimiento *frecuencias*, que se obtienen pulsando consecutivamente el menú *Analizar⇒Estadísticos descriptivos⇒Frecuencias*. El cuadro de diálogo *Frecuencias* se muestra en la figura 5.6. Las variables de la ventana izquierda son todas las variables presentes en el archivo utilizado, mientras que a la derecha están las variables de las que se han solicitado las frecuencias. En este caso, cuyo objetivo es conocer la distribución de todas las variables del archivo Excel, será necesario seleccionarlas todas, cambiando la totalidad de las variables al recuadro de la derecha (con un doble clic de ratón cambian de una ventana a otra). Pulsando el botón *Aceptar* aparecerán las frecuencias dentro de la ventana de resultados (Más adelante, en el capítulo VII, se realizará una exposición pormenorizada del procedimiento frecuencias).



	NUM	PREG12a	PREG12b	PREG13	PREG18	PREG19	PREG20	PREG21
1	1	4	0	1	2	65	4	1
2	2	6	0	2	2	29	4	2
3	3	6	0	1	1	52	0	2
4	4	5	6	2	2	44	5	2
5	5	8	0	4	2	46	2	1
6	6	6	0	3	1	16	4	1
7	7	6	0	4	1	25	0	2
8	8	8	0	1	3	63	1	2
9	9	8	8	2	2	40	2	2
10	10	0	0	4	3	16	0	2
11	11	4	0	1	2	65	4	1
12	12	6	0	2	2	29	4	2
13	13	6	0	1	1	52	0	2
14	14	8	8	1	2	46	2	1
15	15	6	0	4	1	25	0	2
16	16	8	0	3	1	28	2	1
17	17	8	0	1	3	63	1	2

Figura 5.5. Archivo Excel abierto en SPSS.

Figura 5.6. Cuando de diálogo frecuencias.

3. Lectura–recuperación de archivos tipo texto

El programa SPSS permite la recuperación de archivos formato texto al disponer de un asistente para su importación. El menú *Archivo⇒Leer datos de texto* activa el cua-

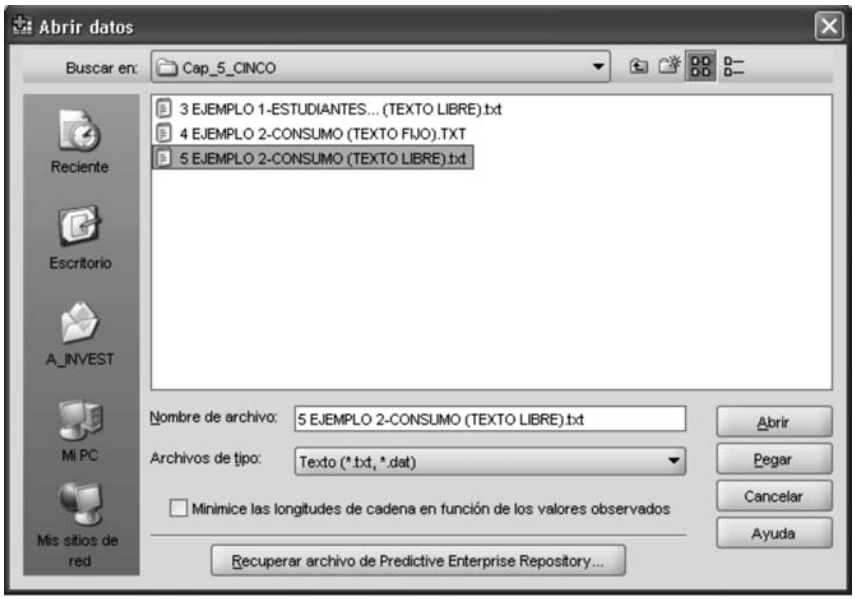


Figura 5.7. Abrir archivo (datos de texto).

dro de diálogo *Abrir archivo* donde se muestran los archivos en formato texto (ver, en la figura 5.7, que en la pestaña "archivos de tipo" aparece automáticamente la extensión texto; *.txt). Tras seleccionar el archivo "Ejemplo 2-consumo (texto libre)" se activa el asistente para la importación de texto, que consta de seis pasos (figura 5.8).

En el primero pregunta si el archivo se ajusta a un formato predefinido, y para poder responder a esta pregunta el programa muestra en la ventana inferior una parte del archivo. Puede visualizarse todo el archivo utilizando las flechas y barras de desplazamiento.

Cuando el archivo no se ajusta a un formato predefinido, que es lo que sucede la mayor parte de las veces, en el segundo paso pregunta cómo están organizadas las variables: delimitadas por un carácter concreto (formato libre), o en columnas de ancho fijo (formato fijo). En la ventana inferior vemos que en este archivo hay un espacio entre las variables, de modo que se trata de un archivo delimitado por un espacio, es decir un archivo en formato libre. Posteriormente se pregunta si se encuentran los nombres en la parte superior del archivo. La vista de la ventana inferior nos lleva a responder negativamente a esta pregunta.

En el paso tres se pregunta por la localización del primer caso (*por defecto* aparece la línea 1), cómo se encuentran representados los casos (por defecto cada línea representa un caso), y el número de casos a importar (todos, por defecto). Posteriormente

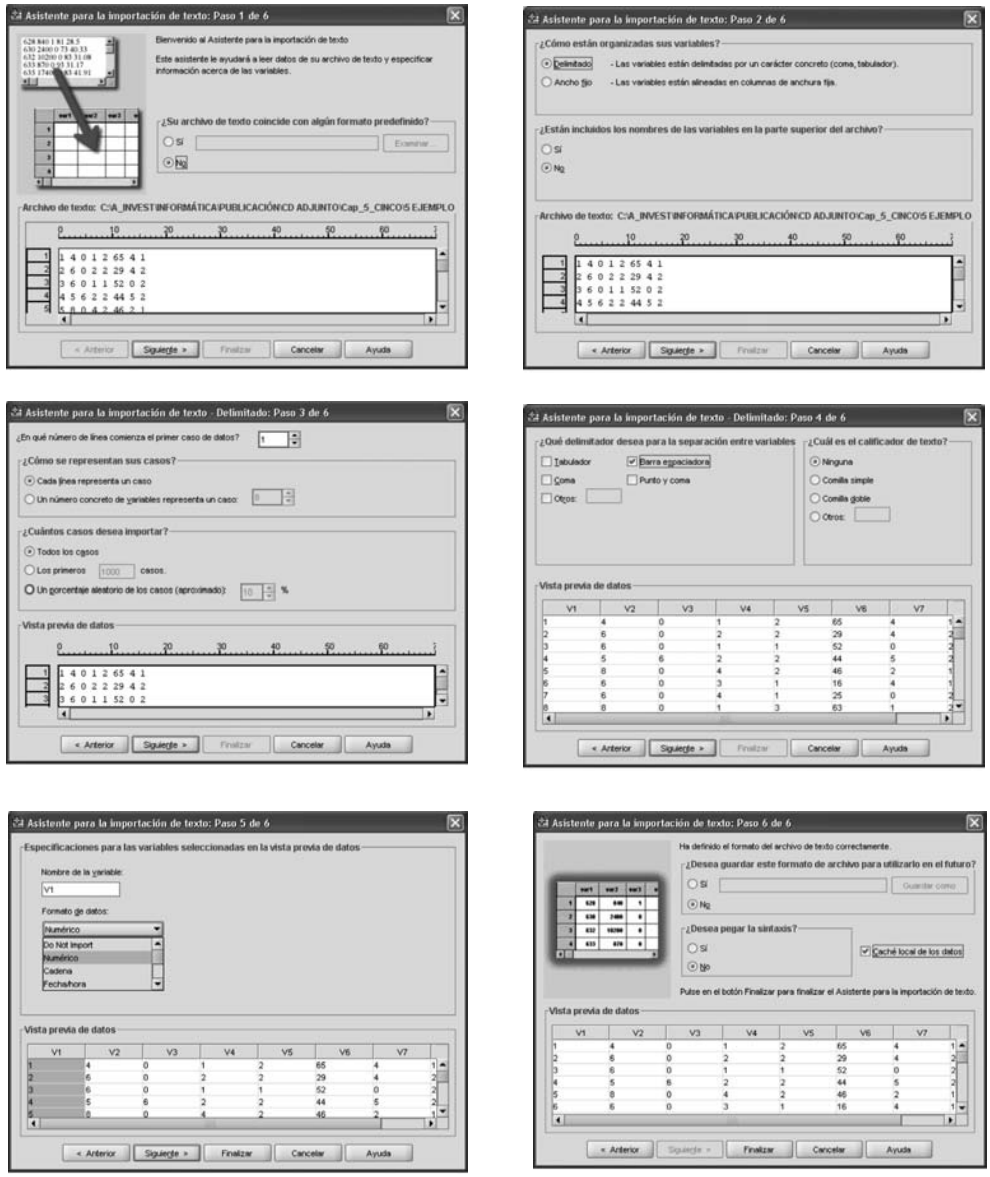
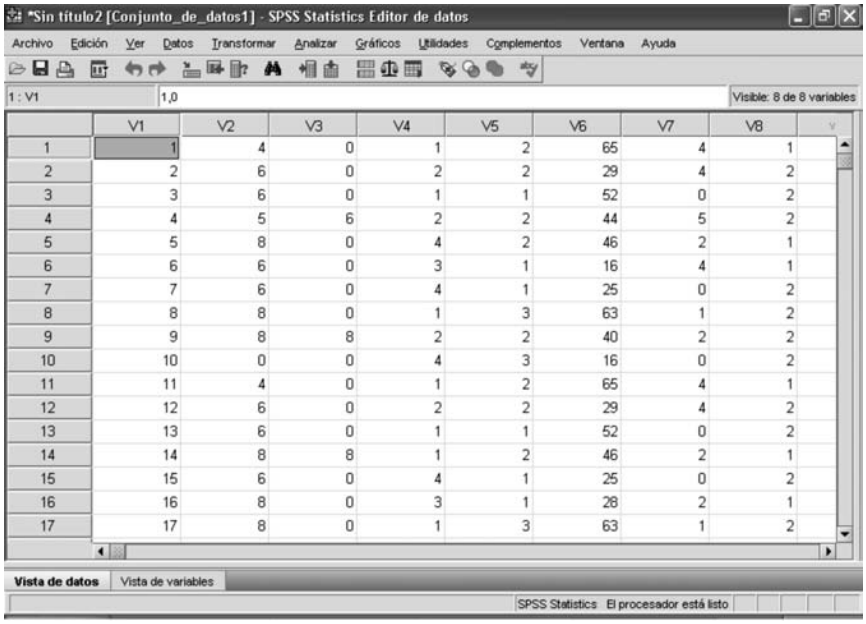


Figura 5.8. Asistente para la importación de texto (formato libre).

(paso cuatro) se pregunta por el carácter que delimita las variables (barra espaciadora), al tiempo que muestra –en la ventana inferior– una vista previa de los datos en el archivo nuevo. Obsérvese como el programa nombra automáticamente cada varia-



	V1	V2	V3	V4	V5	V6	V7	V8
1	1	4	0	1	2	65	4	1
2	2	6	0	2	2	29	4	2
3	3	6	0	1	1	52	0	2
4	4	5	6	2	2	44	5	2
5	5	8	0	4	2	46	2	1
6	6	6	0	3	1	16	4	1
7	7	6	0	4	1	25	0	2
8	8	8	0	1	3	63	1	2
9	9	8	8	2	2	40	2	2
10	10	0	0	4	3	16	0	2
11	11	4	0	1	2	65	4	1
12	12	6	0	2	2	29	4	2
13	13	6	0	1	1	52	0	2
14	14	8	8	1	2	46	2	1
15	15	6	0	4	1	25	0	2
16	16	8	0	3	1	28	2	1
17	17	8	0	1	3	63	1	2

Figura 5.9. Resultado. Archivo texto (formato libre) abierto por SPSS.

ble, con la letra "v" y el número de columna. (No vendrá mal recordar aquí los criterios para la elección del nombre de variables esgrimidos en el apartado 3.2).

En el paso cinco pueden especificarse los nombres y el tipo de las variables. Seleccionada una variable[41] se procede a escribir un nuevo nombre en el espacio correspondiente para, posteriormente, elegir el *formato de los datos*[42]. Tras seleccionar con el ratón la primera variable su nombre aparecerá dentro del recuadro "Nombre de variable", procediendo así variable a variable hasta finalizar con el archivo[43]. No debe pulsarse el botón "*Siguiente>*" hasta haber terminado con el cambio de nombre de todas variables. Si se pulsa por error, lo que dará lugar al sexto cuadro de diálogo de la figura 5.8, siempre podrá volverse al anterior pulsando del botón "*<Anterior*".

Por último el paso seis permite guardar el archivo de recuperación de datos para un uso en un futuro. Se trata del archivo de recuperación, no del archivo de datos, y por esa razón respondemos negativamente a las preguntas formuladas en este cuadro de diálogo. En la figura 5.9 se presenta el resultado del proceso.

41. Poniendo el ratón sobre ella.
42. Dentro de la pestaña "Formato de datos" existe también una opción para no importar una determinada variable (se muestra en la figura 5.8, paso 5).
43. Veremos un ejemplo más adelante.

En la recuperación de la figura 5.9 las variables estaban delimitadas por un espacio, pero ¿qué ocurre cuando se trata de datos sin separación? En la figura 5.10 se muestra la recuperación de un archivo de datos que, con el nombre "EJEMPLO 2-CONSUMO (TEXTO FIJO)", presenta todas las variables sin delimitar, ordenadas en columnas de ancho fijo. En la parte inferior del cuadro de diálogo puede apreciarse como está organizado este archivo.

El proceso de recuperación es prácticamente el mismo que el mostrado en la figura 5.9, y sólo presenta algunas variaciones en el *paso 2* y *paso 4*. En el *paso 2* hay que marcar la opción *Ancho fijo*, y en el "4" el programa solicita que el usuario defina los puntos de división entre variables por medio de líneas verticales. En la parte superior del *paso 4 de 6* (segundo cuadro de diálogo de la figura 5.10) se indica como hacerlo: para insertar una línea basta con hacer un clic de ratón en la parte central de la ventana, donde se muestra el archivo de datos. Estas líneas pueden desplazarse *arrastrándolas* a la posición deseada, y también eliminarse *arrastrándolas* fuera de la ventana de datos. En la segunda ilustración de la figura 5.11 se muestra la delimitación de las variables en el archivo de datos recuperado.

El *quinto paso* permite –como se ha señalado anteriormente– elegir un nombre a cada variable, así como definir el tipo de variable. En este caso cada variable tomará el nombre que se muestra en el libro de códigos incluido dentro de los *materiales complementarios*. Colocado el ratón sobre la primera variable, se pone "num" en el recuadro "nombre de la variable", y se elige el formato de datos correspondiente. Seleccionada la segunda, se procede de la misma forma, y así hasta nombrar todo el archivo. Terminado el proceso de recuperación, el resultado es el mismo que el mostrado en la figura 5.9. A continuación se procede con el etiquetado de las variables.

Antes de finalizar el apartado debemos señalar que los archivos "tipo texto" son los más utilizados en la investigación con encuesta. El Instituto Nacional de Estadística, el CIS, la extinta Fundación CIRES, etc. proporcionan sus archivos de datos en este formato.

4. Lectura–recuperación de archivos de bases de datos

La lectura de archivos de bases de datos es bastante más infrecuente, por el tipo de información incluida en las bases de datos. La recuperación es, también, algo más complicada aunque las últimas versiones de SPSS dispone de un *asistente para bases de datos* que facilita enormemente la lectura de archivos de bases de datos.

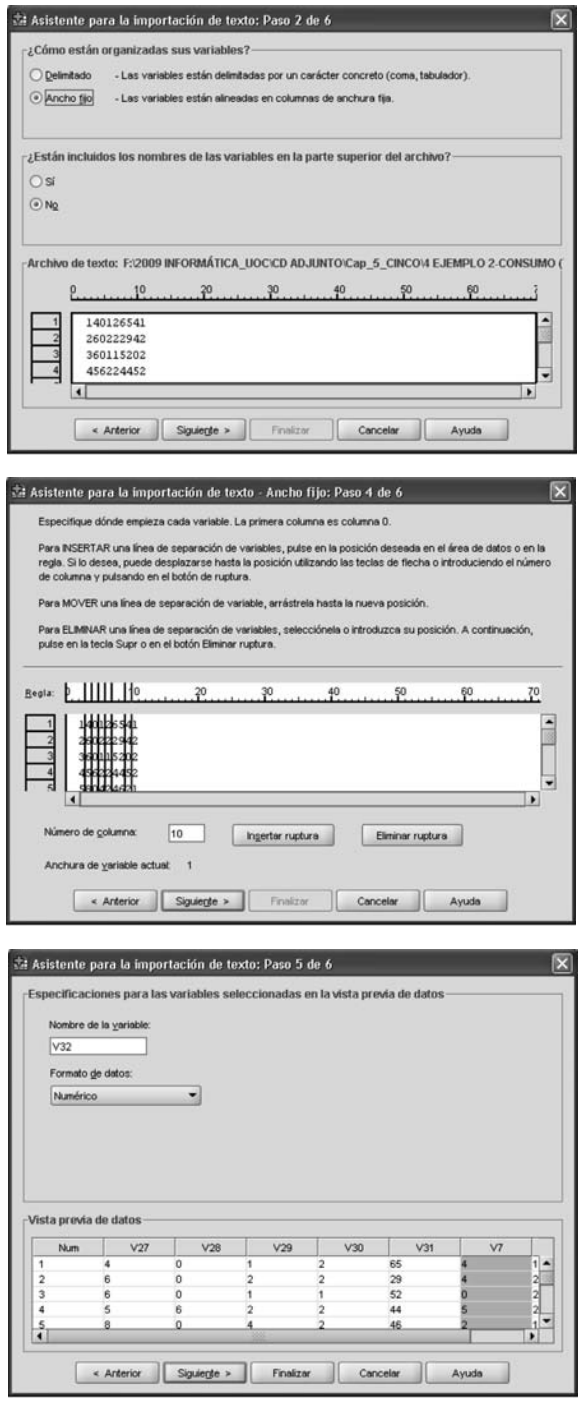
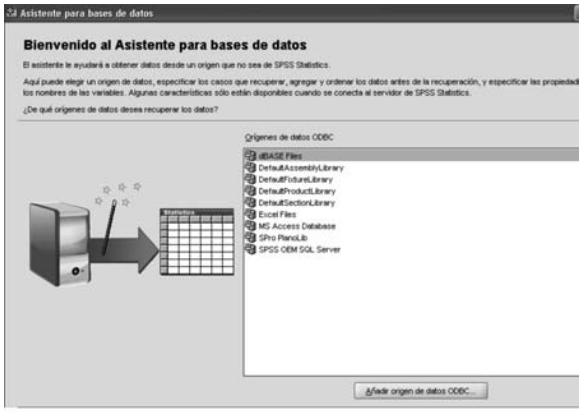
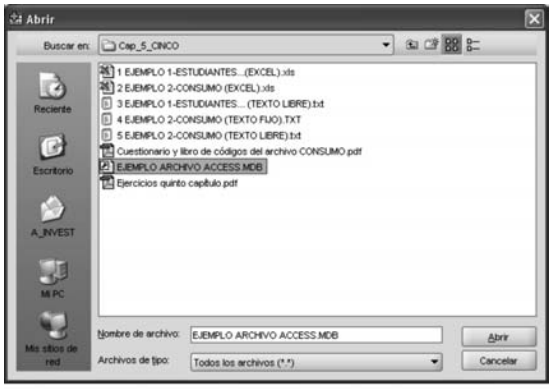


Figura 5.10. Asistente para la importación de texto (formato fijo). Pasos 2, 4 y 5.



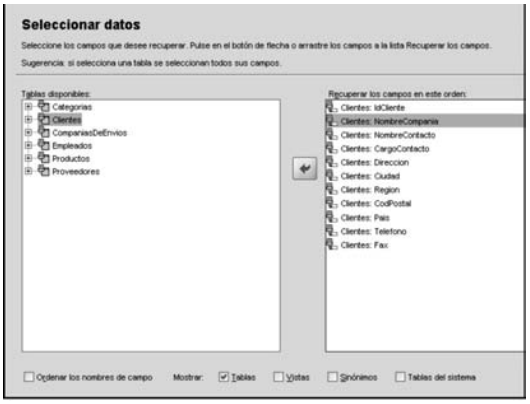
5.11.a

5.11.b



5.11.c

5.11.d



5.11.d

Figura 5.11. Asistente para bases de datos.

	IdCliente	NombreCompania	NombreContacto	CargoContacto	Direcci
1	ALFKI	Alfreds Futterkiste	Maria Anders	Representante de ventas	Obere Str. 57
2	ANATR	Ana Trujillo Emparedados y hela...	Ana Trujillo	Propietario	Avda. de la Const
3	ANTON	Antonio Moreno Taquería	Antonio Moreno	Propietario	Mataderos 2312
4	AROUT	Around the Horn	Thomas Hardy	Representante de ventas	120 Hanover Sq.
5	BERGS	Berglunds snabbköp	Christina Berglund	Administrador de pedidos	Berguvsvägen 8
6	BLAUS	Blauer See Delikatessen	Hanna Moos	Representante de ventas	Forsterstr. 57
7	BLONP	Blondel père et fils	Frédérique Citeaux	Gerente de marketing	24, place Kléber
8	BOLID	Bólido Comidas preparadas	Martín Sommer	Propietario	C/ Araquil, 67
9	BONAP	Bon app'	Laurence Lebihan	Propietario	12, rue des Bouc
10	BOTTM	Bottom-Dollar Markets	Elizabeth Lincoln	Gerente de contabilidad	23 Tsawassen Bl
11	BSBEV	B's Beverages	Victoria Ashworth	Representante de ventas	Fauntleroy Circus
12	CACTU	Cactus Comidas para llevar	Patricio Simpson	Agente de ventas	Cerrito 333
13	CENTC	Centro comercial Moctezuma	Francisco Chang	Gerente de marketing	Sierras de Granad
14	CHOPS	Chop-suey Chinese	Yang Wang	Propietario	Hauptstr. 29
15	COMMI	Comércio Mineiro	Pedro Afonso	Asistente de ventas	Av. dos Lusíadas
16	CONSH	Consolidated Holdings	Elizabeth Brown	Representante de ventas	Berkeley Gardens
17	DRACD	Drachenblut Delikatessen	Sven Ottlieb	Administrador de pedidos	Walserweg 21

Figura 5.12. Archivo Access abierto en SPSS.

Para acceder al asistente hay que abrir el menú *Archivo⇒Abrir base de datos⇒Nueva consulta*, y aparecerá el primero de los cuadros de diálogo de la figura 5.11. El programa pregunta por las fuentes de datos que se desea recuperar la información, mostrando en la ventana de la derecha los formatos disponibles. Esta lista puede ampliarse pulsando el botón *Añadir fuente de datos* (cuadro de diálogo de la figura 5.11.a) En el caso que nos ocupa, recuperar archivos creados por *Microsoft* Access, seleccionaremos *MS Access Database* para pulsar el botón *Siguiente*>.

A continuación aparece el cuadro de diálogo *Acceso al controlador ODBC* (figura 5.11.b), donde se encuentra el botón *Examinar* que –al pulsarlo– abre un cuadro de diálogo con los archivos disponibles en la carpeta de trabajo actual. Será necesario cambiar a la carpeta donde se encuentren los archivos de ejemplos (figura 5.11.c), para seleccionar "EJEMPLO ARCHIVO ACCESS[44]" y, tras pulsar *Abrir*, el SPSS recuperará todos los elementos disponibles en la base de datos (figura 5.11.d). Para elegir los campos a recuperar deben seleccionarse los elementos con los que se desea trabajar y *arrastrarlos* a la ventana de la derecha; donde aparecerán los campos incluidos en cada uno

44. Se trata de un archivo que viene incluido en el programa Microsoft Office con el nombre FPNWIND.mdb. Se le ha cambiado de nombre para facilitar su identificación.

de estos elementos (figura 5.11.e). Pulsando el botón *Finalizar* se obtendrá la figura 5.12, los campos que forman parte de la base de datos "clientes".

De la misma forma que procedimos anteriormente, a continuación será necesario definir las variables para, posteriormente, solicitar las frecuencias con el fin de comprobar que se ha realizado correctamente la recuperación del archivo. No obstante, como la mayor parte de las veces las bases de datos tiene archivos de texto solicitar las frecuencias no permitirá conocer la estructura del archivo recuperado. Por este motivo recomendamos recuperar archivos de bases de datos únicamente cuando en su interior aparezcan códigos numéricos.

Capítulo VI

Depuración de la información

1. Objetivos didácticos del capítulo

Una vez elaborado el archivo de datos es el momento de llevar a cabo una revisión y *depuración* de los datos con el objetivo de evaluar –y si es posible aumentar– la calidad de la información recogida. Se trata de buscar inconsistencias entre ciertas variables, verificar si hay valores que no tienen lugar en determinadas preguntas, analizar las respuestas de las *preguntas filtro*, etc.

A lo largo de este capítulo se expondrán algunas de las estrategias más utilizadas en el proceso de revisión y *depuración* de la información recogida. La primera se fundamenta en la petición de un listado de los valores de todas las variables del cuestionario, realizando tabulaciones para cada variable. Una vez que se ha verificado que todos los valores se ajustan al recorrido de las variables se compara el número de respuestas de las *preguntas filtro* con las *preguntas filtradas*. El siguiente procedimiento consiste en la elaboración de *consistencias lógicas* que deban ser cumplidas por determinadas variables del cuestionario. A continuación se procede con una valoración sobre el nivel de representatividad de las respuestas obtenidas, y posteriormente la ponderación del archivo de datos. Esta fase de revisión y depuración de la información termina con la realización de un primer análisis descriptivo de los datos con el objetivo de conocer los valores presentes en la matriz de datos que muestren *inconsistencias* con el resto de la distribución de los datos.

Como en anteriores ocasiones, utilizaremos los diversos procedimientos de SPSS para resolver cada uno de los objetivos señalados. Para hacer más dinámica la exposición realizaremos la explicación utilizando un archivo de datos sobre una investigación sobre consumo en Navarra cuyo cuestionario –y libro de códigos– se presenta en los *materiales complementarios* (web), dentro de la carpeta *Capítulo 6*[45]. El archivo de datos se ha denominado "Consumo Navarra (No depurado).sav".

45. Debe tenerse en cuenta que se trata de un cuestionario realizado con un encuestador mediante entrevista personal, y eso explica la presencia de una opción –en cada pregunta– para anotar aquellas situaciones en las que el entrevistado "no sabe" o "no contesta". Este hecho explica las diferencias respecto al libro de códigos visto en el apartado 3.9.

2. Listado de valores de las variables

La primera estrategia se fundamenta en pedir un listado de los valores de todas las variables del cuestionario para ver si alguna de ellas tiene valores ajenos a su recorrido (rango): si el sexo del entrevistado (v0038 en el archivo "Consumo Navarra") está codificado como 1 ó 2 no es posible un valor de 7, por ejemplo.

Para realizar esta comprobación es necesario utilizar el procedimiento *frecuencias*, que se obtienen pulsando consecutivamente *Analizar⇒Estadísticos descriptivos⇒Frecuencias*. El cuadro de diálogo *Frecuencias* se muestra en la figura 6.1. Las variables de la ventana izquierda son todas las variables presentes en el archivo utilizado, mientras que a la derecha están las variables de las que solicitaremos las frecuencias[46]. En este caso, cuyo objetivo es conocer los valores fuera de rango de todas las variables del cuestionario, procedemos a seleccionarlas todas, cambiando la totalidad de las variables al recuadro de la derecha (con un doble clic de ratón cambian de una ventana a otra). Pulsando el botón *Aceptar* aparecerán las frecuencias dentro de la ventana de resultados[47].

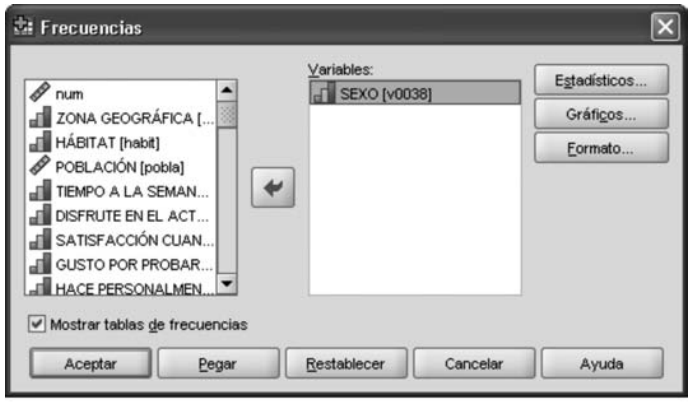


Figura 6.1. Cuando de diálogo frecuencias.

46. Más adelante, en el capítulo VII, se realizará una exposición pormenorizada del procedimiento frecuencias.
47. Aquí tan sólo se mostrará la tabla de frecuencias de una de las variables, aquella en la que se han detectado valores fuera de rango. En el capítulo VII se presentará la ventana de resultados y las posibilidades que ésta ofrece.

En la tabla 6.1 se muestra la *distribución de frecuencias*, correspondiente a la variable v0038, donde llama la atención la presencia de varios "7". ¿Por qué aparece este valor si no fue recogido en el cuestionario? La razón más probable es que se trate de un error en la grabación de datos, de modo que será necesario conocer en qué cuestionarios se ha introducido ese valor.

SEXO

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Hombre	457	50,8	50,8	50,8
	Mujer	434	48,2	48,2	99,0
	7	9	1,0	1,0	100,0
	Total	900	100,0	100,0	

Tabla 6.1. Tabla de frecuencias de la variable sexo del entrevistado (v0038).

Para ello se buscará en el archivo de datos el número de caso donde aparece este valor. Con el SPSS esto se lleva a cabo seleccionando esta variable con un clic de ratón sobre el nombre (v0038), y utilizando después el menú *Edición⇒Buscar*. Será necesario escribir el valor 7 en el espacio dedicado para ese fin (figura 6.2).

Tras efectuar esta tarea el número 7 se localiza en el cuestionario número 17. Localizado el número de caso se buscará el cuestionario original (de papel) para cambiar un valor por el otro. Tras revisarlo se aprecia que se trata de un error en la introducción de datos puesto que en el cuestionario original existe un 2 en esta pregunta

Figura 6.2. Selección de V0038. Buscar datos dentro de esa variable. Menú Edición⇒Buscar.

Ahora bien, podría haber sucedido que en el cuestionario original existiera un 7 en esta variable, si bien es una situación difícil puesto que ha sido revisado por el coordinador de campo y el codificador (como vimos en la sección 2.2 del segundo capítulo). En tal situación se presentan dos opciones: a) definir el 7 como valor perdido y, b) tratar de estimar el sexo del entrevistado a través del resto de variables del cuestionario (*imputación*).

Aprovecharemos esta explicación para presentar algunos aspectos básicos de funcionamiento del SPSS. Uno de los problemas que presentan los cuadros de diálogo de los procedimientos de SPSS es la dificultad para "reconocer" las variables, puesto que tan sólo son visibles las primeras palabras de la etiqueta de la variable; y muchas veces no son suficientes para diferenciar con precisión unas variables de otras (ver, por ejemplo, el cuadro de diálogo de la figura 6.3 y tratar de identificar a que se refiere cada una de las variables que comienzan con "posee...).

Existen dos formas de superar este problema. La primera, y más sencilla, se fundamenta en colocar el puntero del ratón sobre cada una de las variables de la ventana izquierda y, automáticamente, aparecerá el nombre completo de la variable; tal y como puede apreciarse en la figura 6.3. Insistimos en que se trata de colocar sobre la variable el puntero del ratón, que no es necesario pulsar.

Figura 6.3. Cuadro de diálogo Frecuencias.

La segunda es algo más compleja, y consiste en cambiar el formato de visualización de las variables en los cuadros de diálogo. Utilizando el menú *Edición⇒Opciones* aparece el cuadro de diálogo de la figura 6.4. Seleccionada la pestaña *General*, en la parte superior izquierda aparecen varias opciones bajo el título "Listas de variables" referidas a la forma de presentación de las variables en los cuadros de diálogo. Por defec-

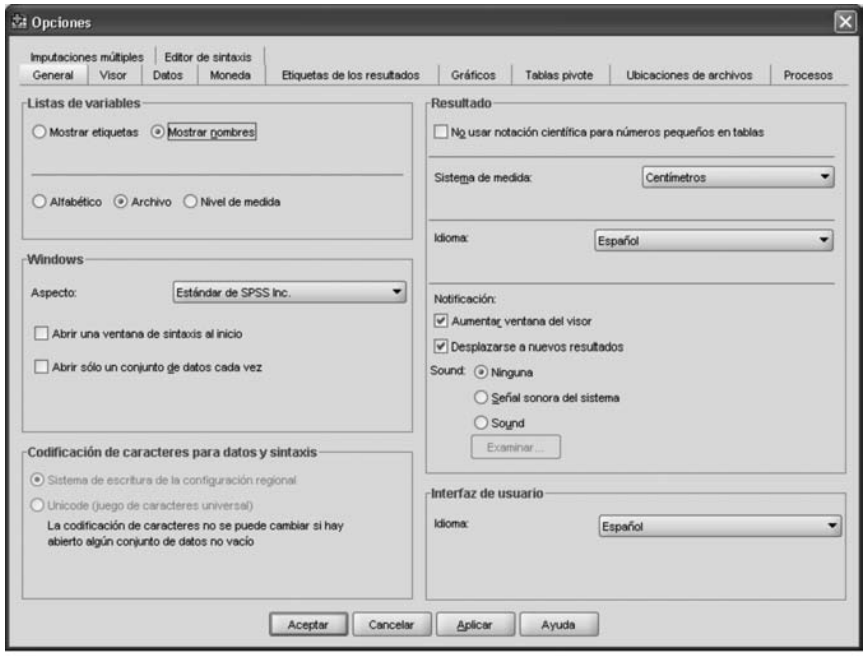


Figura 6.4. Cambio del formato de presentación de las variables en los cuadros de diálogo.

Figura 6.5. Cuadro de diálogo Frecuencias. Lista de variables: mostrar nombres.

to aparece marcado "Mostrar etiquetas", pero ante la dificultad para la localización de las variables marcaremos "Mostrar nombres". Tras pulsar el botón *Aceptar* el programa advierte que el cambio en la visualización de las listas de variables surtirá efecto la próxima vez que se abra un archivo de datos.

A partir de este momento en los cuadros de diálogo sólo aparece el nombre de la variable (v0001, v0005, v0008, como se muestra en la figura 6.5), de modo que es preciso tener cerca el libro de códigos o el cuestionario[48].

Existe otra alternativa, que consiste en definir las etiquetas de las variables incluyendo el nombre de la variable, de modo que se muestra conjuntamente el nombre de la variable junto con su etiqueta de identificación. Así hemos procedido en el archivo "Encuestas estudiantes...".

Expuesta la forma de visualización de las variables, seguimos con la depuración de variables, dejando que sea el lector el que opte por la situación que le parezca más apropiada.

3. Preguntas filtro y preguntas filtradas

Una vez constatado que todos los valores se ajustan al recorrido de las variables será interesante comparar el número de respuestas de las *preguntas filtro* y las *preguntas filtradas*. Seleccionemos como ejemplo las variables v005 y v006 del archivo "Consumo Navarra". La primera de éstas pregunta si suele hacer personalmente las compras de ropa y calzado, mientras que la segunda se interesa por quién compra la ropa, y debe ser respondida únicamente por aquellos que "alguna vez", "rara vez" y "nunca" realizan sus compras de ropa y calzado[49].

Si 162 personas compran su ropa "alguna vez", "rara vez" y "nunca", (54, 61 y 47 respectivamente), es evidente que no puede haber más de 162 respuestas a la pre-

48. En realidad el investigador debe tener a mano *en todo momento* el cuestionario o el libro de cogidos; aún cuando estén todas las variables correctamente identificadas y etiquetadas.

49. Texto de ambas preguntas:

Preg. 05. ¿SUELE HACER PERSONALMENTE (Sólo o acompañado) LA COMPRA DE ROPA Y CALZADO PARA USTED?

– Siempre	1	Preg. 7
– Casi siempre	2	Preg. 7
– Alguna vez	3	Preg. 6
– Rara vez	4	Preg. 6
– Nunca	5	Preg. 6
– NS/NC	0	

*[SOLO PARA LOS QUE **NO** COMPRAN LA ROPA "SIEMPRE" Y "CASI SIEMPRE"]*

Preg. 06. ¿QUIEN LE SUELE COMPRA LA ROPA?

– Madre	1
– Marido/mujer/compañera	2
– Otros	3
– NS/NC	0

gunta siguiente. Sin embargo la V006 (pregunta 6) ha sido respondida por 166 personas, es decir hay cuatro personas que no debían responder esta pregunta. En este caso surge el problema de cuál de las dos respuestas será la correcta. ¿La solución?, volver al cuestionario original para ver si ha sido un problema en la grabación de la información. A continuación se detalla como llevar a cabo tal comprobación:

1. Frecuencias de las variables v005 y v006.

HACE PERSONALMENTE LA COMPRA DE ROPA Y CALZADO (V005)

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	NS/NC	6	,7	,7	,7
	Siempre	589	65,4	65,4	66,1
	Casi siempre	143	15,9	15,9	82,0
	Alguna vez	54	6,0	6,0	88,0
	Rara vez	61	6,8	6,8	94,8
	Nunca	47	5,2	5,2	100,0
	Total	900	100,0	100,0	

¿QUIEN SUELE COMPRA LA ROPA? (V006)

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Madre	28	3,1	16,9	16,9
	Marido, mujer, compañero /a	127	14,1	76,5	93,4
	Otros	11	1,2	6,6	100,0
	Total	**166**	18,4	100,0	
Perdidos	No procede	734	81,6		
Total		900	100,0		

Tabla 6.2. Tabla de frecuencias de las variables v005 y v006.

2. La v006 únicamente debe ser respondida por aquellos que "alguna vez", "rara vez" y "nunca" compran su propia ropa, los que señalan las opciones 3, 4 y 5 en v005: 54, 61 y 47 (tabla 6.2). En total, 162 personas señalan –en v005– que no realizan sus propias copras de ropa y calzado.
3. Pero v006 es respondida por 166 personas, hay 4 entrevistados que han respondido v006 pero que no deberían hacerlo porque no han dicho –en v005– que "alguna vez", "rara vez" y "nunca" compran su propia ropa.
4. Detección del problema mediante la ordenación del archivo de datos:
   - Menú *Datos*⇒Ordenar casos (Figura 6.6), colocando v005 (hace la compra de ropa y calzado para usted) en la ventana de la derecha. Orden de clasificación: Ascendente.

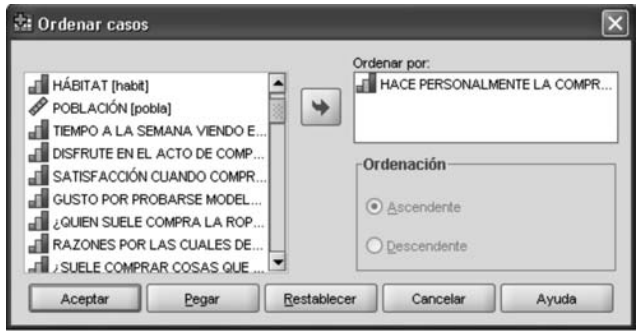


Figura 6.6. Ordenación de casos según la variable v005.

   - El archivo de datos aparecerá ordenado según la frecuencia con la que hacen las compras de ropa y calzado (figura 6.7).
   - En las líneas 3, 4, 5, y 6 se aprecian varios entrevistados que no han respondido v005 (la no respuesta se ha codificado con el valor 0) y si lo han hecho en v006.
   - A la izquierda (variable "num") puede observar el número de encuesta de estos valores, que corresponden a los cuestionarios número 136, 142, 609 y 614.
5. De nuevo se volverán a revisar estos cuestionarios. Es difícil que una pregunta de este tipo no haya sido detectada por el coordinador de campo, pero puede suceder.
6. Cuando el origen del problema está en la recogida de datos (es decir, que los cuestionarios están mal respondidos) lo más conveniente será asignar los valores de ambas variables (v005 y v006) como perdidos, puesto que desconocemos cual de los dos es el erróneo. Es decir, ¿no han respondido v005 porque no les han preguntado?, ¿porque no sabían como hacerlo, etc...? ¿Por qué motivo han respondido v006, si únicamente debe ser respondida por los que responden los valores 3, 4 ó 5 en v005?

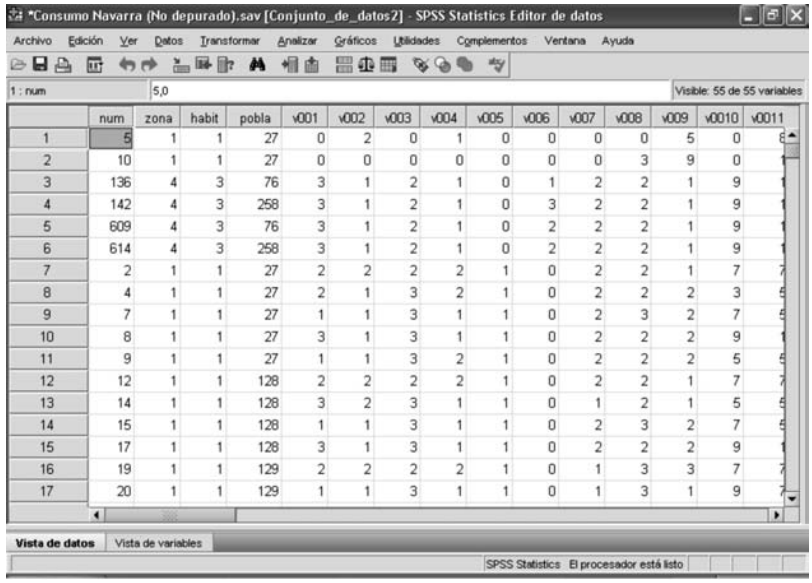


Figura 6.7. Editor de datos ordenado casos según la variable v005.

7. Ante la imposibilidad de indicar al programa que asigne como perdidos determinados valores de cuatro cuestionarios, la forma más rápida de hacerlo es cambiando estos valores por otros no utilizados. Por ejemplo cambiar los "0" de v005 por el valor 50 en los cuestionarios afectados (números 136, 142, 609 y 614). En v006 el 1 (del cuestionario 136) puede sustituirse por 51, el 3 (del cuestionario 142) por 53 y los 2 (cuestionarios 609 y 612) por 52. Estos valores deberán definirse como perdidos.
8. A continuación se define cada uno de éstos como valor perdido, tal y como se muestra en la figura 6.8.

Figura 6.8. Valores perdidos de v005.

4. Comprobación de *consistencias lógicas* entre variables

Otra estrategia de depuración consiste en la elaboración de consistencias lógicas o *edits* que deban ser cumplidos por los datos del cuestionario, con el objetivo de localizar los casos que no las cumplen. Recordemos la situación comentada anteriormente (apartado 2.2 del capítulo II) donde una persona con 8 hijos declaraba tener 15 años. Conviene realizar la elaboración de estas reglas de consistencias lógicas en el proceso de construcción del cuestionario, puesto que es el momento más adecuado para poder definir las respuestas contradictorias. Además es necesario realizar un detallado análisis de los *edits* antes de utilizarlos en el archivo de datos a fin de detectar inconsistencias entre éstos.

La elaboración de consistencias lógicas en este archivo comienza observando que 163 entrevistados[50] responden la V0028 (actividad del cónyuge), lo que lleva a preguntarnos si estarán todos casados o viviendo en pareja:

1. Seleccionamos V0035 igual a 1 (solteros), para conocer si alguna persona soltera (o sin pareja) ha respondido a la pregunta "actividad del cónyuge".
   Para ello utilizaremos el procedimiento Seleccionar casos (*Datos⇒Seleccionar Casos*). En la figura 6.9 marcar "Si se satisface la condición" y, tras pulsar el botón *Si la op...*, aparecerá el cuadro de diálogo de la figura 6.10 donde se escribe la condición lógica a cumplir (v0035=1). Haciendo un clic de ratón sobre v0035, y pulsando esta variable se desplaza a la ventana de la derecha. A continuación bastará con marcar los operandos de la parte inferior; concretamente el "=" y "1". Más adelante, en el capítulo VIII, analizaremos con más detalle el procedimiento *Seleccionar casos*.
   Pulsando *Continuar* se vuelve al cuadro de diálogo de la figura 6.9 (obsérvese que en la parte inferior aparece "Descartar casos no seleccionados"), y con *Aceptar* se llevará a cabo la selección.
2. Frecuencias de "Actividad del cónyuge" (v0028). Al analizar estos resultados se aprecia que hay tres personas solteras que responden a esta pregunta (tabla 6.3). Como puede verse en el cuestionario incluido en los *materiales complementarios*, los cónyuges que no trabajan no deben responder la v0028, asignándoles el valor "0".
3. Será necesario analizar específicamente estos tres casos. Para ello recomendamos seleccionar V0035 igual a 1 (solteros), y V0028 (actividad del cónyuge) distinto de 0 (cónyuges que trabajan), como se muestra en la figura 6.11[51].

---

50. Menú *Analizar⇒Estadísticos Descriptivos⇒Frecuencias⇒*v0028.
51. Recuperando el cuado de diálogo anterior, figura 6.10, bastará con añadir el operando "&" y escribir la segunda condición lógica.

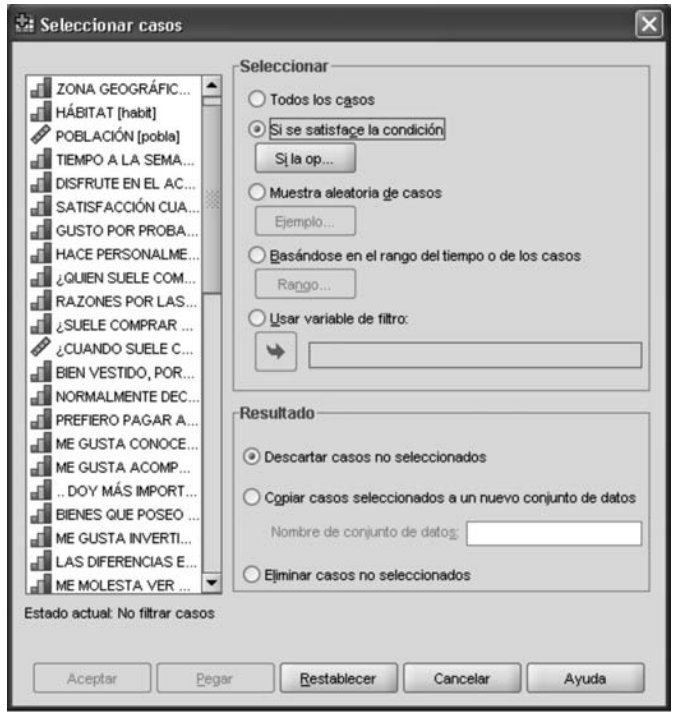


Figura 6.9. Selección de casos.

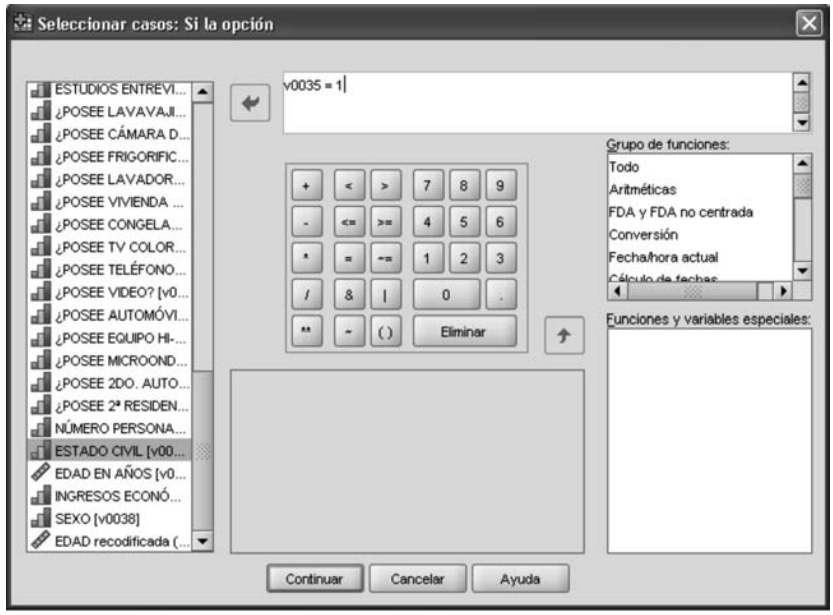


Figura 6.10. Selección de casos: condición lógica.

ACTIVIDAD CONYUJE					
	Frecuencia	**Porcentaje**	**Porcentaje válido**	**Porcentaje acumulado**	
Válidos	STATUS. OCUP. ALTO	1	,3	33,3	33,3
	STATUS. MEDIO	1	,3	33,3	66,7
	STATUS BAJO	1	,3	33,3	100,0
	Total	3	,8	100,0	
Perdidos	0	360	99,2		
Total	363	100,0			

Tabla 6.3. Selección de casos: condición lógica.

Ahora bien, aunque el procedimiento *Seleccionar casos* debe utilizarse siempre marcando la opción "Descartar casos no seleccionados" (figura 6.9), en este momento marcamos "E*liminar* casos no seleccionados" para detectar los entrevistados que no cumplen la situación especificada.

4. El resultado, que se muestra en la figura 6.12, desvela los tres casos que no cumplen la condición especificada; concretamente los cuestionarios 359, 810 y 860. Se trata de hombres (v0038=1), cabezas de familia (v0029=1), solteros (v0035=1), y que responden a la pregunta sobre la profesión del cónyuge (v0028). Ante esta situación solo cabe una solución: revisar los cuestionarios y, en caso de que el error persista, eliminar estos sujetos de la muestra. Sin duda los lectores son ahora más conscientes de la importancia de la revisión de los cuestionarios durante la realización de los trabajos de campo.
   En este caso, afortunadamente, se trataba de un error a la hora de grabar las respuestas de los cuestionarios al archivo de datos. La persona encargada de realizar esta tarea había introducido tres valores erróneos en la variable v0028).
5. Muy importante. Cuando se termine la sesión *no guardar* el archivo de datos resultante, puesto que se han eliminado la mayor parte de los casos[52]. Tan sólo interesa anotar la situación de la figura 6.12 para hacer los cambios en el archi-

52. Con todo, y por lo que pudiera pasar, siempre es recomendable guardar una copia de seguridad del archivo de datos fuera del ordenador.

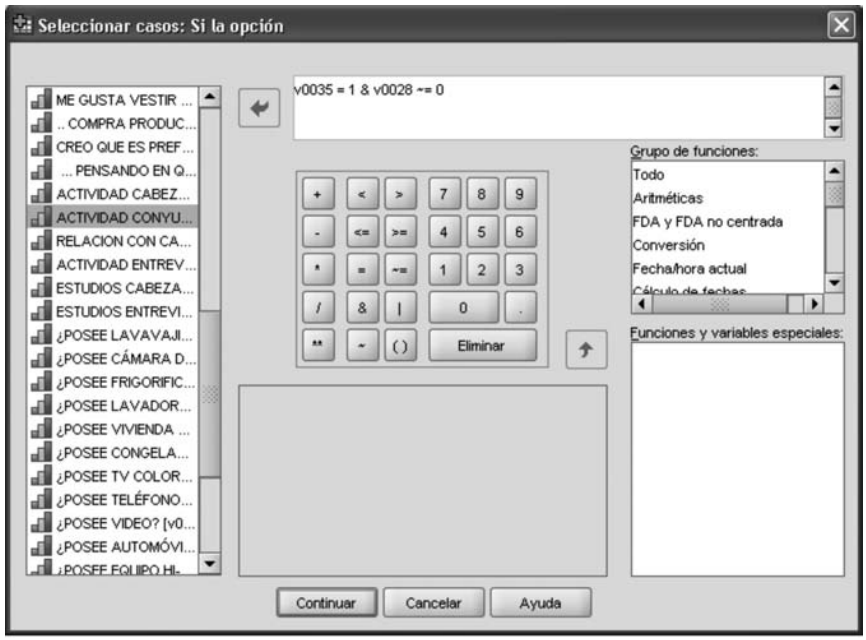


Figura 6.11. Selección de casos: condición lógica con dos operandos.

	num	v0027	v0028	v0029	v0035	v0036	v0037	v0038
1	359	6	6	1	1	27	5	1
2	810	6	8	1	1	27	5	1
3	860	6	4	1	1	27	5	1
4								
5								

Figura 6.12. Resultado de la selección.

vo original. Una vez recuperado el archivo bastará con cambiar –en la variable V028– el valor 6 del cuestionario 359 por un 0, el 8 del cuestionario 810 por un 0, y el 4 del 860 por otro 0 (recordar que en este ejemplo la no respuesta está codificada ceros, como puede apreciarse en el cuestionario incluido en los *materiales complementarios*).

5. Nivel de representatividad de las respuestas obtenidas

El proceso de validación de las variables estará condicionado al nivel de representatividad de las respuestas, o dicho de otra forma, por el número de respuestas obtenidas en cada variable. Cuando una pregunta tiene una alta tasa de no respuesta (definida según el ratio número de no respuestas/muestra total) se puede optar por presentar los cuadros y tablas de datos con estos valores, o bien eliminarlos de los análisis al definirlos como valores perdidos. Veamos una aplicación con el archivo de datos utilizado en este capítulo (consumo Navarra.sav):

1. Detección de "no respuestas" en cada variable. Solicitando las frecuencias de todo el archivo (o volviendo a las frecuencias utilizadas en el apartado 6.2), y recordando que "no responde" fue codificado con el valor 0, detectamos el bajo número de "no respuestas".
2. Definir como valor perdido el código utilizado para la definición de las preguntas filtro (si procede[53]).
3. Cuantificación –y asignación como valores perdidos– de los códigos que indican otros aspectos como "más de una respuesta"[54]. Aunque no todas las variables presentan dobles respuestas, se ha eliminado esta opción en la figura 6.13 a modo de ejemplo de cómo proceder con las variables que se cumplen estas tres situaciones con los códigos del archivo "Encuestas estudiantes...".

Figura 6.13. Valores perdidos (no respuestas, dobles respuestas y preguntas filtro).

53. En el archivo *Consumo Navarra* no se diferencia entre "no respuesta por filtro" y "no respuesta por otros factores" (es decir, se ha empleado el mismo código para ambas situaciones); situación contraria al archivo utilizado en Encuestas Estudiantes... Téngase en cuenta que en este último se ha operado de forma diferente al codificarse la no respuesta por filtro con el valor 90 y otras no respuestas con el 99.
54. Si procede (ver nota anterior).

Tras definir los valores perdidos se solicitan las frecuencias de cada variable[55] a fin de comprobar las modificaciones efectuadas. El resultado se muestra en la tabla 6.4 que, como podemos comprobar, difiere en algunos aspectos de la tabla 6.1 (frecuencias del sexo sin eliminar valores perdidos). Un análisis comparativo entre ambas figuras desvela que las columnas "Frecuencia" y "Porcentaje" son similares; y que difieren significativamente en las dos columnas de la derecha. En la tabla 6.4 se han calculado los porcentajes eliminando aquellos valores definidos como perdidos; esto es, los que fueron introducidos incorrectamente por el codificador[56].

SEXO

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Hombre	457	50,8	51,3	51,3
	Mujer	434	48,2	48,7	100,0
	Total	891	99,0	100,0	
Perdidos	7	9	1,0		
	Total	900	100,0		

Tabla 6.4. Tabla de frecuencias del sexo (v0038).

El problema de la no respuesta aumenta cuando la tasa de cooperación es menor: si una pregunta ha sido contestada únicamente por la mitad de la población el error muestral se incrementará, reduciéndose con ello la precisión de las estimaciones. Pese a la importancia de este problema, a nuestro juicio surge otro mucho mayor al cuestionarnos si los que han contestado pueden ser considerados como una muestra representativa de los que no lo han hecho. Dicho de otro modo, nos preguntamos si los que contestan son iguales que los que no lo hacen. Si los que responden tuvieran el mismo perfil, mismas opiniones, iguales características, etc. que los que no lo hacen podríamos decir que unos representan a los otros; el problema es que numerosas investigaciones han señalado importantes diferencias entre los que responden a todas las preguntas y los que no lo hacen: los que más responden suelen tener un mayor nivel educativo, un status socioeconómico elevado, más mujeres que hombres, jóvenes y residentes en hábitat urbano. Existen dos estrategias para solucionar este pro-

55. Menú *Analizar*⇒*Estadísticos Descriptivos*⇒*Frecuencias*⇒v0038.
56. En el apartado 2 se explicó cómo debe procederse con estos valores. Tan sólo deben definirse como valores perdidos cuando no es posible aplicar otra solución.

blema: la más sencilla es dejar la pregunta como está, asumiendo que los que no contestan son iguales a los que contestan. La otra es realizar una *imputación* de las no respuestas.

Además del análisis pregunta a pregunta, conviene considerar también el promedio de preguntas *no contestadas* por cada entrevistado a fin de localizar y conocer las características de los entrevistados con altas tasas de no respuesta. Si un sujeto tiene un elevado número de no respuestas la mejor solución será eliminarlo del archivo de datos.

Decíamos que el archivo de datos está formado por filas en las que están introducidos los sujetos y por columnas donde están las variables. ¿Como hacer para obtener la información de cada sujeto en las distintas variables? Para realizar este proceso con el programa SPSS será necesario cambiar la *posición* del fichero de datos considerando las variables como sujetos y los sujetos como variables. Esta operación se realiza con el Menú *Datos⇒Transponer*, seleccionando a continuación todas las variables del cuestionario. Posteriormente bastará con examinar los datos de las nuevas variables, que ahora serán sujetos puesto que el fichero ha sido "movido" noventa grados. Veamos, de nuevo, una aplicación con el archivo de datos utilizado como ejemplo:

1. Menú *Datos⇒Transponer*, pasando todas las variables al recuadro de la derecha (figura 6.14).

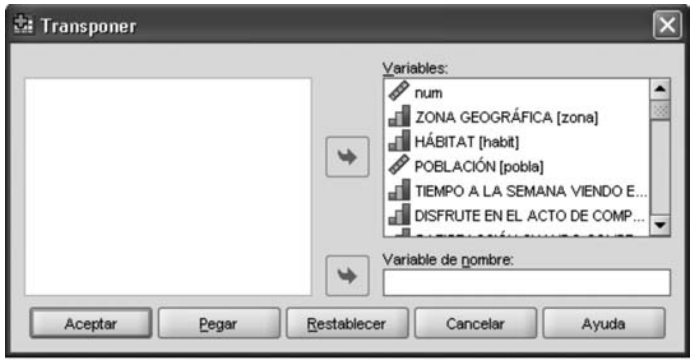
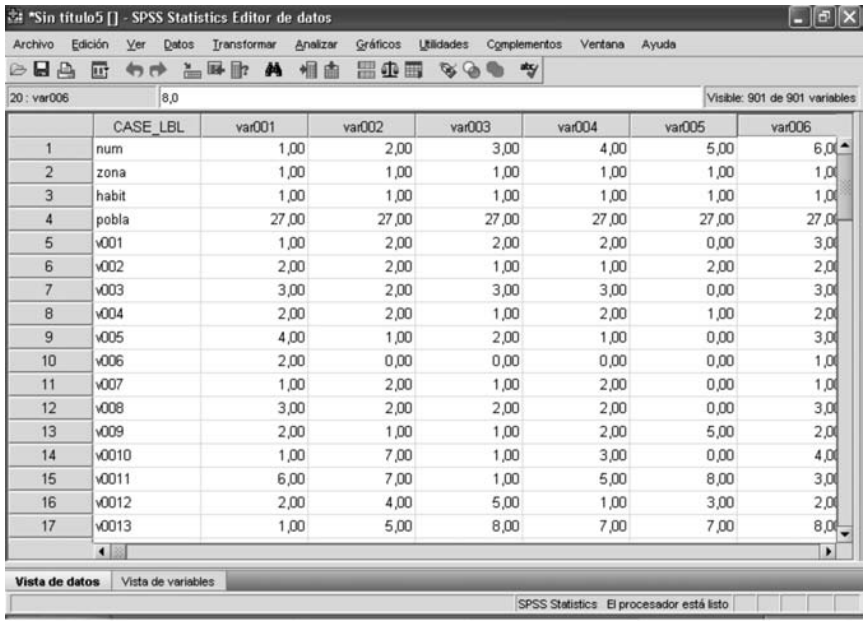


Figura 6.14. Transponer archivo (Menú Datos⇒Transponer).

2. En este momento debe tenerse en cuenta que en columnas aparecen los sujetos entrevistados, y en filas cada una de las variables. De hecho, el programa crea automáticamente una nueva variable, llamada CASE_LBL, que contiene los nombres originales de cada variable (figura 6.15).
   Para comprender mejor el efecto de este procedimiento recomendamos comparar el archivo obtenido con el *original*, que fue mostrado en la figura 6.7 (pági-



	CASE_LBL	var001	var002	var003	var004	var005	var006
1	num	1,00	2,00	3,00	4,00	5,00	6,0
2	zona	1,00	1,00	1,00	1,00	1,00	1,0
3	habit	1,00	1,00	1,00	1,00	1,00	1,0
4	pobla	27,00	27,00	27,00	27,00	27,00	27,0
5	v001	1,00	2,00	2,00	2,00	0,00	3,0
6	v002	2,00	2,00	1,00	1,00	2,00	2,0
7	v003	3,00	2,00	3,00	3,00	0,00	3,0
8	v004	2,00	2,00	1,00	2,00	1,00	2,0
9	v005	4,00	1,00	2,00	1,00	0,00	3,0
10	v006	2,00	0,00	0,00	0,00	0,00	1,0
11	v007	1,00	2,00	1,00	2,00	0,00	1,0
12	v008	3,00	2,00	2,00	2,00	0,00	3,0
13	v009	2,00	1,00	1,00	2,00	5,00	2,0
14	v0010	1,00	7,00	1,00	3,00	0,00	4,0
15	v0011	6,00	7,00	1,00	5,00	8,00	3,0
16	v0012	2,00	4,00	5,00	1,00	3,00	2,0
17	v0013	1,00	5,00	8,00	7,00	7,00	8,0

Figura 6.15. Resultado de la *transposición*.

na 141). Ahí se aprecia con claridad que en la primera columna aparece el número de encuesta (con un orden no correlativo puesto que aquel archivo había sido ordenado considerando las respuestas de V005), en la variable segunda la zona, después el hábitat, la población (ver, por ejemplo, el gran número de entrevistas realizadas en el municipio 27), y a continuación cada una de las variables del cuestionario. Esta información aparece en *filas* dentro del archivo transpuesto: la primera recoge el número de encuesta, la segunda fila la zona, a continuación el hábitat, la población (véase la "repetición" del número 27 detectado anteriormente) y, por último, cada una de las variables del cuestionario.

3. Definir, en cada columna, los valores perdidos pertinentes: 0 en el ejemplo archivo que estamos utilizando; 98 y 99 en el ejemplo "Encuestas estudiantes...".
4. Solicitar las frecuencias de todas las *variables*, teniendo en cuenta el total que aparece en la línea antepenúltima de la tabla de frecuencias.

   En la tabla 6.5 se muestran las frecuencias de var004, que corresponden al cuarto entrevistado. El número total de preguntas del cuestionario son 54 (última fila de la tabla), y este entrevistado ha dejado una sin responder (valor 0), de modo que presenta una tasa de respuesta del 98,1% (total de la fila antepenúltima).

var004					
		Frecuencia	**Porcentaje**	**Porcentaje válido**	**Porcentaje acumulado**
Válidos	1,00	14	25,9	26,4	26,4
	2,00	18	33,3	34,0	60,4
	3,00	3	5,6	5,7	66,0
	4,00	2	3,7	3,8	69,8
	5,00	4	7,4	7,5	77,4
	6,00	5	9,3	9,4	86,8
	7,00	2	3,7	3,8	90,6
	8,00	3	5,6	5,7	96,2
	27,00	1	1,9	1,9	98,1
	44,00	1	1,9	1,9	100,0
	Total	53	98,1	100,0	
Perdidos	,00	1	1,9		
Total		54	100,0		

Tabla 6.5. Frecuencias de una variable transpuesta; de un entrevistado.

5. Eliminar, si procede, los sujetos con alto número de preguntas no respondidas; por ejemplo "var005" (tabla 6.6). En la tabla 6.6 se muestra la tabla de frecuencias de un entrevistado, concretamente el número 5, que únicamente ha respondido 28 preguntas del cuestionario, es decir, tan sólo el 51,9% de las preguntas. Ante esta situación, lo más correcto es eliminar este caso del archivo de datos; si bien esta situación debiera haberse detectado antes de grabar las respuestas en el archivo de datos (de nuevo apelamos a la necesidad que una exhaustiva revisión de los cuestionarios durante la realización del trabajo de campo).
   Debe tenerse muy presente que esta forma de proceder puede reducir representatividad muestral. Ahora bien, ¿es factible una investigación basada en una gran muestra de entrevistados que no responden a la mayor parte de las preguntas

var005					
		Frecuencia	**Porcentaje**	**Porcentaje válido**	**Porcentaje acumulado**
VVálidos	1,00	7	13,0	25,0	25,0
	2,00	7	13,0	25,0	50,0
	3,00	1	1,9	3,6	53,6
	4,00	1	1,9	3,6	57,1
	5,00	3	5,6	10,7	67,9
	6,00	1	1,9	3,6	71,4
	7,00	2	3,7	7,1	78,6
	8,00	4	7,4	14,3	92,9
	27,00	1	1,9	3,6	96,4
	46,00	1	1,9	3,6	100,0
	Total	28	51,9	100,0	
Perdidos	,00	26	48,1		
Total		54	100,0		

Tabla 6.6. Frecuencias de un entrevistado (variable transpuesta) con elevada tasa de no respuesta.

del cuestionario? La experiencia investigadora nos dice que poco podemos hacer con estos entrevistados.

La eliminación deberá llevarse a cabo en el archivo original, no en el traspuesto. Para ello cerramos el archivo traspuesto (una vez anotados los sujetos a eliminar), abrimos "Consumo Navarra" y procedemos con la eliminación del entrevistado número 5. Existen dos formas de llevar a cabo esta eliminación:

a. La primera consiste en seleccionar ese determinado sujeto, mediante un clic de ratón en el margen izquierdo de la ventana de datos (concretamente en la parte sombreada), y posteriormente pulsar –en el teclado– la tecla *suprimir*.

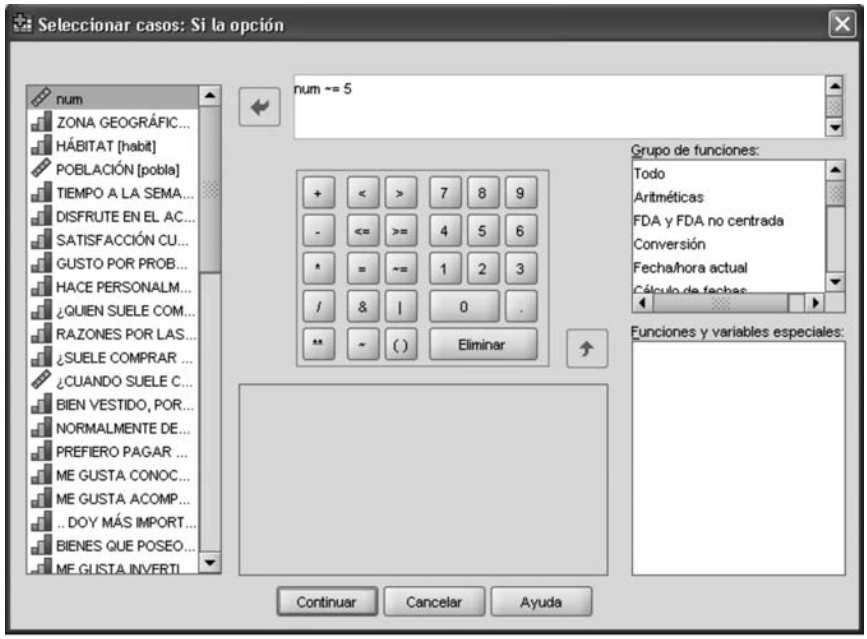
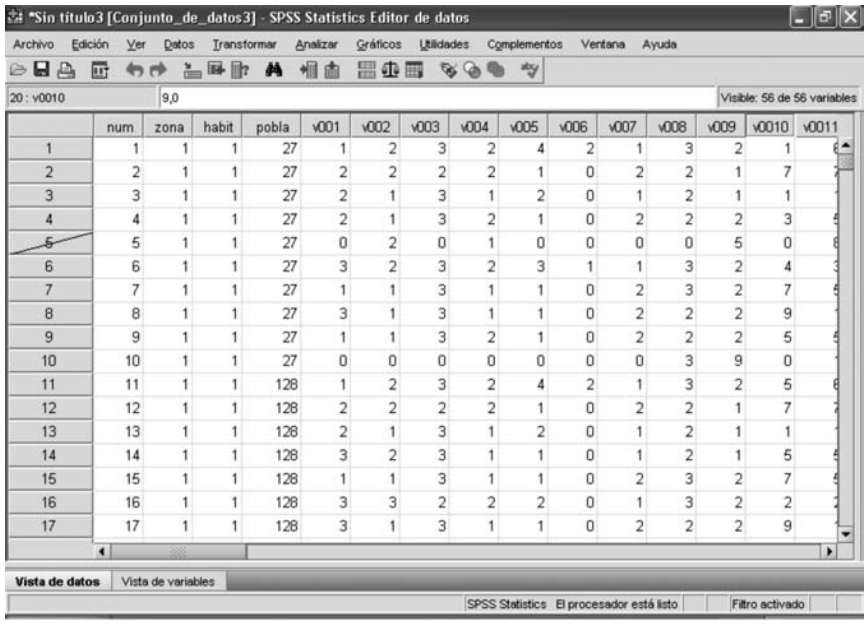


Figura 6.16. Selección de casos: eliminación de un caso.



	num	zona	habit	pobla	v001	v002	v003	v004	v005	v006	v007	v008	v009	v0010	v0011
1	1	1	1	27	1	2	3	2	4	2	1	3	2	1	6
2	2	1	1	27	2	2	2	2	1	0	2	2	1	7	7
3	3	1	1	27	2	1	3	1	2	0	1	2	1	1	1
4	4	1	1	27	2	1	3	2	1	0	2	2	2	3	5
5	5	1	1	27	0	2	0	1	0	0	0	0	5	0	6
6	6	1	1	27	3	2	3	2	3	1	1	3	2	4	3
7	7	1	1	27	1	1	3	1	1	0	2	3	2	7	5
8	8	1	1	27	3	1	3	1	1	0	2	2	2	9	1
9	9	1	1	27	1	1	3	2	1	0	2	2	2	5	5
10	10	1	1	27	0	0	0	0	0	0	0	3	9	0	1
11	11	1	1	128	1	2	3	2	4	2	1	3	2	5	6
12	12	1	1	128	2	2	2	2	1	0	2	2	1	7	7
13	13	1	1	128	2	1	3	1	2	0	1	2	1	1	1
14	14	1	1	128	3	2	3	1	1	0	1	2	1	5	5
15	15	1	1	128	1	1	3	1	1	0	2	3	2	7	5
16	16	1	1	128	3	3	2	2	2	0	1	3	2	2	2
17	17	1	1	128	3	1	3	1	1	0	2	2	2	9	1

Figura 6.17. Editor de datos con un caso no seleccionado (caso 5).

b. La segunda utiliza el procedimiento seleccionar casos (Datos⇒Seleccionar Casos); para elegir después la variable Num (número de cuestionario) e indicarle al programa que seleccione todos los casos excepto el 5 (ver figura 6.16). A partir de este momento el caso número 5 aparecerá con una barra a la izquierda, que indica que esta des-seleccionado (figura 6.17), que no será considerado en los análisis posteriores.
Recomendamos esta forma de proceder, en vez de la anterior, puesto que se conserva el valor original; pudiéndolo recuperar cuando sea preciso.

6. Ponderar archivo

En numerosas ocasiones la información recogida mediante encuestas muestrales se desvía ligeramente de las condiciones planificadas en el diseño muestral, de modo que es necesario equilibrar los datos recogidos utilizando ponderaciones. El fin es *devolver* a cada estrato la proporcionalidad que tiene en la realidad de donde ha sido extraída la muestra. En la tabla 6.7 se expone la muestra teórica[57], la muestra obtenida, y los coeficientes de ponderación.

	Muestra Teórica	**Muestra obtenida**	**Coeficientes ponderación**
16-25 años	225 (25,0%)	211 (23,4%)	1,068376
26-35 años	190 (21,1%)	206 (22,9%)	0,921397
36-45 años	174 (19,3%)	174 (19,3%)	1,000000
46-55 años	156 (17,3%)	155 (17,2%)	1,005814
56-65 años	155 (17,2%)	154 (17,1%)	1,005848
TOTAL	900	900	

Tabla 6.7. Muestra teórica, muestra obtenida (real), coeficientes de ponderación.

57. Información obtenida con el procedimiento frecuencias (Analizar⇒Estadísticos descriptivos⇒Frecuencias⇒v0036_re). La variable v0036_re es la edad (v0036) en grupos: 16–25 años, 26–35 años, 36–45 años, 46–55 años, 56–65 años. En el capítulo VIII se explicará el procedimiento utilizado para llevar a cabo esta transformación.

El proceso de ponderación comienza con el cálculo de los coeficientes de ponderación, obtenidos dividiendo la muestra teórica entre la muestra real: estos coeficientes se recogen en la cuarta columna de la tabla 6.7, y se calculan dividiendo la distribución porcentual de muestra teórica entre la obtenida: 25,0/23,4=1.074; 21,1/22,9=0.92,…). Posteriormente será necesario introducir estos valores en el archivo de datos activo, en una nueva variable que nombraremos –en este caso– con el nombre de *Pond*[58]. Existen varias formas de introducir los coeficientes de ponderación en esta variable:

a. Directamente, teniendo en cuenta el valor de la edad en la variable v0036.
b. Ordenando el archivo según la edad del entrevistado (el procedimiento "ordenar casos" se ha mostrado en la figura 6.6), para introducir posteriormente los coeficientes de ponderación.
c. Utilizando el procedimiento *seleccionar casos* (figura 6.9 y 6.10), seleccionar cada una de las situaciones apuntadas (edad < 25; edad entre 26 y 35, etc.), e introducir posteriormente cada valor en las celdillas seleccionadas.
d. Utilizando el procedimiento "IF" en el editor de sintaxis[59]. Instrucción:
   - IF (V0036=1) POND = 1,068376.
   - IF (V0036=2) POND = 0,921397.
   - IF (V0036=3) POND = 1,000000.
   - IF (V0036=4) POND = 1,005814.
   - IF (V0036=5) POND = 1,005848.
   - EXECUTE.
e. Empleando el procedimiento *Recodificar en distintas variables*[60] con POND como "variable de resultado". Ejemplo: Variable de entrada v0036 (edad), variable salida = pond. Pasos a seguir:
   1. Valores antiguos de 16 a 25, valores nuevos = 1,071423.
   2. Valores antiguos de 26 a 35, valores nuevos = 0,926830.
   3. Valores antiguos de 36 a 45, valores nuevos = 1,000000.
   4. Valores antiguos de 46 a 55...
   5. Aceptar

Una vez que se ha construido la variable con la ponderación bastará con seleccionar del menú principal la opción *Datos⇒Ponderar casos* (figura 6.18). Posteriormente será necesario marcar la opción "*Ponderar casos mediante*" y, haciendo doble clic sobre

58. En este momento conviene introducir esta variable en el libro de códigos.
59. El editor de sintaxis se explica en el próximo capítulo, sección 2.
60. Este procedimiento se explica en el capítulo ocho, apartado 5. Tras explicar ese procedimiento se presenta un ejercicio sobre cómo realizar esta recodificación.

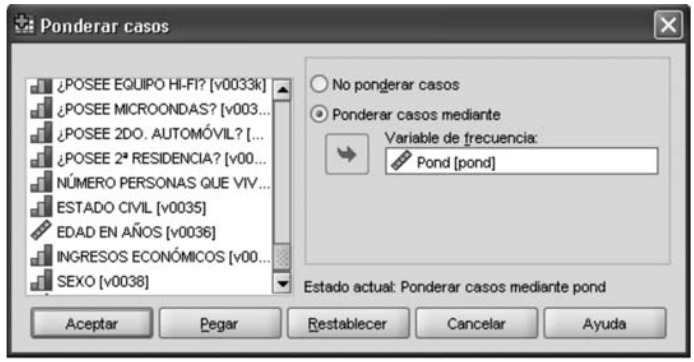


Figura 6.18. Ponderar archivo.

la variable *Pond*, ésta pasará al recuadro de la derecha. Pulsando el botón *Aceptar* la muestra quedará ponderada. A partir de este momento aparecerá la palabra *Ponderado* en la barra de estado, situada en la esquina inferior derecha del Editor de datos. En la tabla 6.8 se muestran las frecuencias de la variable v0036_re con la muestra ponderada y sin ponderar.

	Muestra sin ponderar	**Muestra ponderada**
16-25 años	211 (23,4%)	226 (25,1%)
16-25 años	206 (22,9%)	191 (21,2%)
36-45 años	174 (19,3%)	174 (19,3%)
46-55 años	155 (17,2%)	155 (17,2%)
56-65 años	154 (17,1%)	154 (17,1%)
TOTAL	900	900

Tabla 6.8. Resultados de v0036_re: muestra ponderada y sin ponderar.

PARTE III

ANÁLISIS DE LA INFORMACIÓN RECOGIDA

Capítulo VII

La obtención de información

1. Objetivos didácticos del capítulo

Tras la depuración del archivo de datos comienza la apasionante aventura de analizar las respuestas de los entrevistados. En este capítulo se muestran los procedimientos más utilizados para realizar el primer análisis de los datos presentes en el editor de datos de SPSS: *Frecuencias*, *Frecuencias de respuestas múltiples*, y *Descriptivos*.

La utilización de cada uno dependerá del tipo de datos a analizar. Aunque el procedimiento *Frecuencias* puede utilizarse para cualquier tipo de datos, normalmente se emplea para el análisis de datos cualitativos, para escalas nominales y ordinales. Las *Frecuencias de respuestas múltiples* precisan del mismo tipo de datos, si bien su empleo está recomendado cuando se analizan preguntas con más de una respuesta (*preguntas multirrespuesta*). Preguntas como: ¿cuáles son las *dos* situaciones que mejor definen tu actividad en tu tiempo libre (pregunta 3 del cuestionario presentado en el apartado 6 del capítulo II); ¿qué asignaturas que proponen libros de lectura obligatoria (pregunta 7); ¿qué periféricos o dispositivos tienen los ordenadores de tu hogar (pregunta 17a); etc. Se trata de unas preguntas muy utilizadas en la investigación con encuesta y con un tratamiento ligeramente diferente al resto de preguntas del cuestionario.

Por último el procedimiento *Descriptivos* se emplea con escalas de intervalo o razón (variables cuantitativas, numéricas), y proporciona estadísticos de resumen de las variables distribuidos en cuatro grandes grupos: a) Medidas de tendencia central; b) medidas de dispersión (*desviación típica*, varianza y rango); c) medidas de la forma de la distribución (asimetría y curtosis); y d) otras medidas como el valor mínimo, valor máximo y la suma de valores.

Durante todo el texto la explicación se llevará a cabo utilizando el archivo de datos "ENCUESTAS ESTUDIANTES 2002_03.SAV", elaborado en el tercer capítulo; de modo que será necesario abrir este archivo para seguir los ejemplos. Para recuperar un archivo de datos hay que seleccionar *Archivo⇒Abrir⇒Datos*, o pulsar el símbolo de la barra de herramientas.

2. Frecuencias de variables nominales y ordinales

Seleccionando, del editor de datos, el menú *Analizar* aparecen las distintas opciones de análisis de datos, que varían según las partes del programa adquiridas. La selección de cada opción da lugar a distintas técnicas de análisis de datos. Para explicar el funcionamiento de los procedimientos estadísticos nos centraremos en la opción *Estadísticos Descriptivos* y, dentro de ésta, seleccionaremos *Frecuencias* a fin de exponer la forma básica de funcionamiento con el SPSS.

Tras seleccionar consecutivamente *Analizar⇒Estadísticos descriptivos ⇒Frecuencias* se obtiene una pantalla con dos ventanas y con un *botón* con una *flecha* en el centro que permitirá desplazar las variables de una ventana a otra, tal y como se muestra en la figura 7.1. En la ventana de la izquierda aparecen todas las variables presentes en el archivo abierto, mientras que en la ventana de la derecha –titulada con el nombre *Variables*– están aquellas con las que se desean realizar los análisis pertinentes, las distribuciones de frecuencias en este caso. En un primer momento la ventana derecha aparece vacía de modo que será preciso seleccionar las variables pertinentes y pasarlas a esta ventana. Existen varias formas de realizar esta operación:

Figura 7.1. Cuadro de diálogo *Frecuencias*.

- Haciendo doble clic sobre la variable (esta opción se vio en el capítulo anterior, cuando utilizamos el procedimiento frecuencias en la depuración de datos)
- Señalar una a una con el ratón, y pulsar el *botón-flecha* del centro cada vez.

- Cuando el interés es seleccionar variables consecutivas deberemos marcar la primera y arrastrar el ratón hasta la última[61]. Después se pulsará el *botón-flecha* entre ventanas.
- Otra opción, cuando interesen variables alternas, es tener pulsada la techa *control* al seleccionar cada variable con el ratón. Una vez que se han seleccionado aquellas que interesan basta con pulsar el *botón-flecha* del centro y todas pasarán de una a otra ventana.

En caso de seleccionar por error una variable que no interesa bastará con realizar el proceso contrario: al marcar con el ratón las variables de la ventana derecha la *flecha* automáticamente cambiará de sentido y, al hacer clic sobre ella las variables pasarán a la ventana izquierda.

Antes de continuar es preciso realizar un pequeño comentario sobre los botones de acciones que aparecen en la parte inferior, y que se repiten en todos los procedimientos estadísticos del SPSS. De derecha a izquierda:

- *Ayuda*: ayuda del proceso estadístico elegido.
- *Cancelar*: cierra la pantalla de diálogo, cancelando todos los cambios realizados desde que se abrió por última vez.
- *Restablecer*: limpia la pantalla de selección de variables y vuelve las opciones seleccionadas a la posición original de partida. Debe tenerse en cuenta que al utilizar por segunda vez un cuadro de diálogo este conserva las órdenes de la utilización anterior.
- *Aceptar*: ejecuta el proceso estadístico correspondiente.
- *Pegar*: la selección realizada por el interfaz gráfico de Windows –seleccionando distintos elementos del cuadro del diálogo– queda convertida en una frase de instrucciones de *sintaxis* de SPSS.
  Un clic en el botón *Pegar* convierte la selección realizada por la interfaz gráfica de Windows en una frase formada por los *comandos* y *subcomandos* que eran utilizados por las versiones para MS-DOS del SPSS. Esta frase se escribe en una ventana de texto (Figura 7.2) que aparece automáticamente tras pulsar el botón *Pegar*. Todo el texto incluido en esta ventana puede ser guardado y recuperarse en sesiones posteriores. Para grabarlo hay que seleccionar el menú *Archivo⇒ Guardar como* (en la ventana de sintaxis) y escribir el nombre elegido. El programa añadirá la extensión "sps".

61. Otra posibilidad es, una vez marcada la primera variable, marcar la última pulsando (a la vez) la tecla Shift.

Es preciso señalar que pese a la sustancial mejora en la facilidad de funcionamiento de las versiones Windows del SPSS, con el interfaz de Windows no es posible acceder a todas las operaciones del programa[62]. Funcionar con la ventana de sintaxis permite, además de introducir nuevos elementos no disponibles en los menús del SPSS, guardar todos los mandatos realizados en una sesión de trabajo para volverlos a utilizar más tarde. Desde nuestro punto de vista la gran aportación de esta forma de funcionar es evitar la repetición de una larga secuencia de cuadros de diálogos cuando necesitamos realizar muchas operaciones repetitivas.

La Figura 7.2 muestra los componentes incluidos en la ventana de sintaxis. En la parte superior aparece la lista de funciones del menú principal, con un nuevo menú (*Ejecutar*) para ejecutar los comandos de sintaxis. Esta función permite ejecutar toda la ventana de sintaxis, o una selección determinada.

Figura 7.2. Editor de Sintaxis de SPSS.

Figura 7.3. Barra de herramientas del Editor de Sintaxis.

En la figura 7.3 se muestra la *Barra de herramientas* del editor de sintaxis, que presenta muchas similitudes a la barra del Editor de datos explicada en la sección 4.4. Los aspectos diferentes, que serán los que explicaremos aquí, aparecen situados en la parte derecha:

62. No se asuste el lector puesto que en la mayoría de los casos es suficiente con la utilización del interfaz gráfico. En el anexo 1 se muestran las órdenes –utilizando el lenguaje de sintaxis de SPSS– de los análisis realizados en el capítulo.

- Ir a datos.
- Ejecuta la orden de sintaxis.
- Continuar ejecución de sintaxis.
- Utilizar conjuntos de variables.
- Mostrar todas las variables.
- Proporciona ayuda de sintaxis sobre la técnica estadística en pantalla.
- Crear/editar autoproceso.
- Ejecutar proceso.
- Designar ventana.

Finalizaremos este apartado exponiendo una estrategia de funcionamiento para los usuarios iniciados en las versiones Windows del SPSS y que tengan curiosidad por conocer las operaciones que pueden ser realizadas en la ventana de sintaxis. Esta estrategia se fundamenta en, una vez elaborada la instrucción mediante los cuadros de diálogo, *pegarla* en el editor de sintaxis y acudir al editor para modificar únicamente los elementos precisos. Unas páginas atrás justificábamos la elección de este programa basados, entre otras razones por la accesibilidad y claridad de los menús de ayuda. Es hora de probar tal elección. Tras *Pegar* la orden *Frecuencias* realizada mediante el interfaz de Windows (figura 7.2), estaremos interesados en conocer nuevas posibilidades de este procedimiento. Pulsando con el ratón el icono aparecerá una ventana de ayuda con todas las opciones disponibles en el procedimiento *Frecuencias* (figura 7.4).

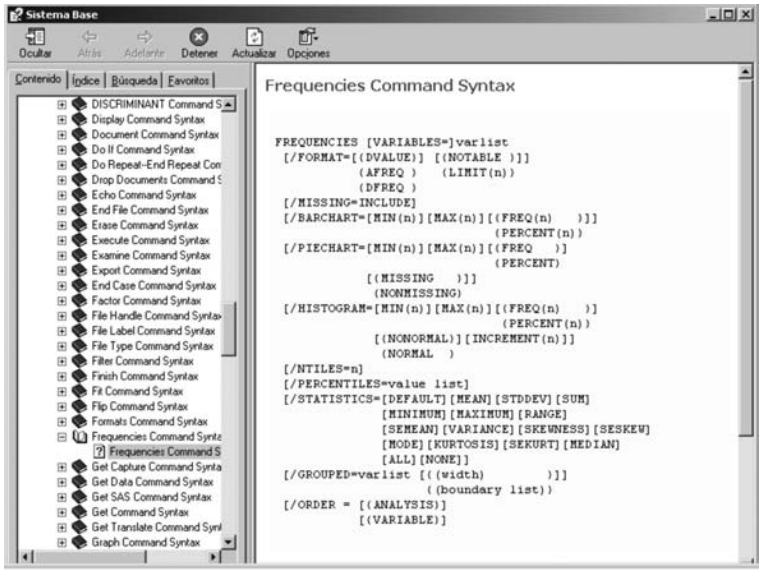


Figura 7.4. Ayuda de sintaxis: instrucciones del procedimiento *Frecuencias*.

Terminada la explicación de los botones de *acciones* de SPSS comunes a todos los procedimientos estadísticos, volvamos al cuadro de diálogo de la figura 7.1. Antes de proceder con el cálculo y visionado de la tabla de frecuencias, pulsando el botón *Aceptar* o ejecutando la orden de la ventana de sintaxis, volveremos a esa figura para explicar otros cuatro botones específicos del procedimiento *Frecuencias*. Son los que aparecen en la parte derecha de la Figura 7.1, y que dan lugar a los cuadros de diálogo secundarios. El primero, situado en la parte superior, hace referencia a los estadísticos a calcular en la variable seleccionada. Al pulsar este botón aparece un cuadro de diálogo (Figura 7.5) con 15 medidas agrupadas en cuatro grupos:

- Valores percentiles, referido a los índices de posición: *cuartiles*, deciles y percentiles.
- Medidas de dispersión: desviación típica, varianza, amplitud (*rango*), valor mínimo, valor máximo y error típico de la media
- Medidas de tendencia central: *media aritmética*, *mediana*, *moda* y la suma de valores.
- Análisis de la distribución de los datos: asimetría y curtosis.

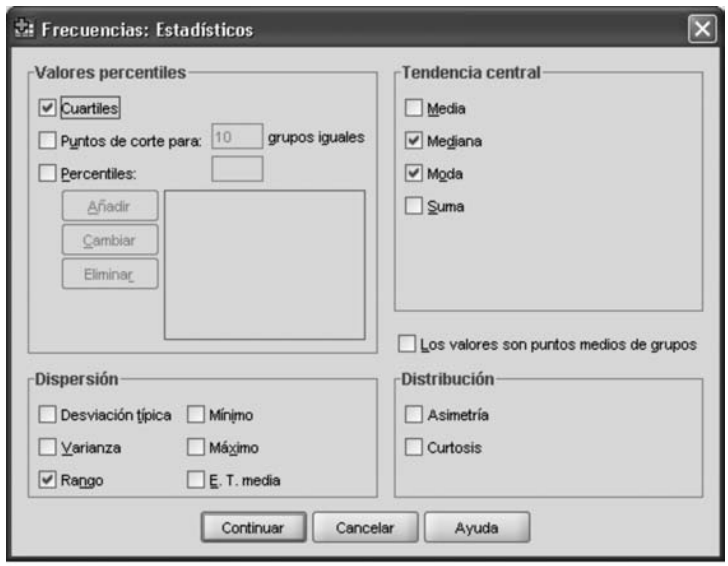


Figura 7.5. Cuadro de diálogo *Frecuencias*: Estadísticos.

En caso de no estar interesados en ninguno de ellos, pulsando el botón *Cancelar* se vuelve a la pantalla anterior. Aquí hemos seleccionado los cuartiles, la mediana, moda

y rango. Tras realizar esta selección, pulsando el botón *Continuar* se vuelve al menú principal del procedimiento *Frecuencias* (Figura 7.1).

El segundo botón específico del procedimiento Frecuencias está referido a la presentación gráfica de los resultados. Tras seleccionar el botón *Gráficos* del menú principal *Frecuencias* aparece el cuadro de diálogo de la figura 7.6, que realiza gráficos de barras, de sectores, e histogramas. El Histograma, que permite sobre-escribir la curva normal, se realiza siempre con porcentajes, mientras que los gráficos de barras y sectores posibilitan la utilización de datos relativos o porcentajes.

Figura 7.6. Cuadro de diálogo *Frecuencias*: Gráficos.

Es importante señalar aquí una instrucción que afecta a todos los cuadros de diálogo de SPSS. Siempre que se muestren diversas opciones señaladas con círculos –como sucede en el cuadro de diálogo de la figura 7.6– sólo se podrá elegir una. En el tema que nos afecta habrá que seleccionar un gráfico, o ninguno seleccionando la primera línea, y un tipo de valores: frecuencias o porcentajes. Tras elegir, por ejemplo, Gráficos de Barras con porcentajes, pulsaremos *Continuar* para volver al menú principal del procedimiento *Frecuencias*.

Tan sólo resta explicar un botón, el situado más abajo, que se refiere al *formato* de los resultados, y que afecta a la ordenación de la *distribución de frecuencias* y al tratamiento de varias variables (cuando se solicita más de una). En este caso dejamos las opciones por defecto, tal y como se muestra en la figura 7.7, y pulsando *Continuar* se vuelve al menú principal del procedimiento *Frecuencias*.

Antes de pulsar *Aceptar* puede ser interesante ver cómo los estadísticos y los gráficos han modificado el menú de sintaxis del procedimiento frecuencias. Pulsando el botón *Pegar* aparece la figura 7.8 que, como, vemos difiere significativamente de la

Figura 7.7. Cuadro de diálogo *Frecuencias*: Formato.

orden presentada en el editor de sintaxis de la figura 7.2. Las dos primeras líneas y la última son iguales en ambas figuras. Respecto a las diferencias, en la tercera línea de la figura 7.8 aparece la solicitud de cuartiles, en la cuarta el rango, mediana y moda, y en la quinta el gráfico de barras (*barchart*) con porcentajes.

Figura 7.8. Editor de sintaxis con los procedimientos de las figuras 7.1, 7.5, 7.6 y 7.7.

Pulsando el icono *Ejecutar* ( ) del editor de Sintaxis (figura 7.8) el programa procederá a ejecutar el procedimiento seleccionado con todas sus opciones. Otra forma, más habitual, es pulsar el botón *Aceptar* del procedimiento frecuencias de la figura 7.1. Al análisis y comentario de los resultados obtenidos dedicaremos el siguiente apartado.

Por último, es preciso indicar que pulsando con el botón secundario del ratón sobre cualquier elemento de los cuadros de diálogo, el programa muestra información sobre cada uno de los aspectos explicados.

3. Resultados de SPSS: visor de resultados y editor de gráficos

Comenzaremos con el visor de resultados. En la parte superior izquierda aparece el icono de resultados de SPSS, símbolo que estará presente en todos los archivos de resultados. A la derecha se encuentra el nombre del archivo resultados, en este caso (figura 7.9) el utilizado por defecto: *Resultado1*. Para cambiarlo basta con seleccionar el menú *Archivo⇒Guardar como* y añadir un nuevo nombre. El icono de resultados de SPSS y el nombre del archivo aparece también en la parte inferior de la pantalla, en la barra de tareas.

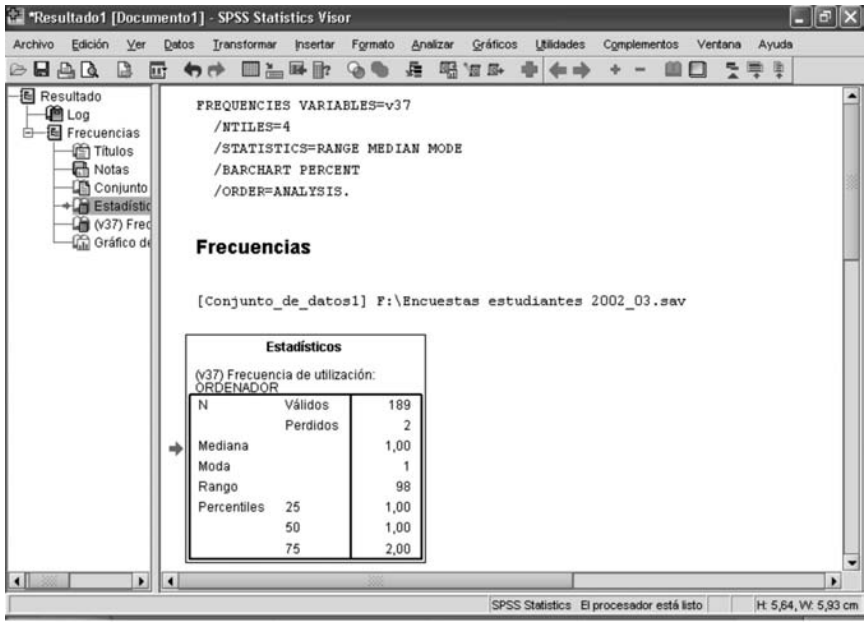


Figura 7.9. Visor de resultados de SPSS.

En la segunda línea está situada la barra de menús, con una disposición muy similar a la mostrada en el Editor de datos (figura 3.1), si bien aquí aparecen dos nuevos los menús *Insertar* y *Formato*. Este último permite alinear los resultados a la izquierda, a la derecha o al centro; mientras que *Insertar* incluye un conjunto de acciones como Insertar o eliminar salto de página, insertar nuevo encabezado, nuevo título o nuevo texto, insertar diversos tipos de gráficos, así como un archivo de texto y un objeto (figura 7.10).

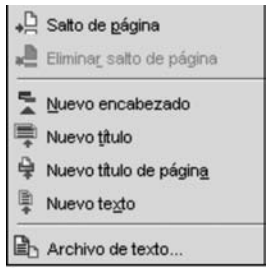


Figura 7.10a. Visor de resultados: menú *Insertar*.

Figura 7.10b. Visor de resultados: menú *Formato*.

En las líneas siguientes aparecen las *barras de herramientas*, con iconos que permiten acceder rápidamente a los procedimientos más usuales de trabajo, sin necesidad de acudir al menú y a los diferentes submenús. Procederemos con su explicación, de izquierda a derecha:

- Abrir archivo de resultados.
- Guardar archivo de resultados.
- Imprimir resultados, permitiendo la selección de una parte.
- Presentación preliminar.
- Exportar resultados.
- Rellamada mostrando las últimas operaciones realizadas con el programa.
- Deshacer y rehacer.
- Activa la ventana de edición de datos.
- Activa la ventana de edición de datos y permite ir a un caso.
- Activa la ventana de edición de datos y permite ir a una variable.
- Ofrece información sobre variables.
- Utilizar conjuntos de variables.
- Mostrar todas las variables.
- Selecciona los últimos resultados.
- Asociar autoproceso.
- Crear/editar autoproceso.
- Ejecutar proceso.
- Designar ventana.
- Ascender.
- Descender.
- Expandir elementos de titulares seleccionados.
- Contraer.
- Mostrar elementos seleccionados.
- Ocultar elementos seleccionados.

Insertar encabezado.
Nuevo título.
Nuevo texto.

Por último la ventana con los resultados, que está dividida en dos partes, tal y como puede apreciarse en la figura 7.9. A la derecha se muestran los resultados solicitados[63] con sus correspondientes tablas, estadísticos, gráficos, etc.; mientras que la parte izquierda presenta un esquema-índice con los titulares de todos los resultados contenidos en la parte derecha, lo que posibilita acceder muy rápidamente a la búsqueda y lectura de éstos. Es posible cambiar las dimensiones de cada una situándose sobre la barra de separación, pulsando el botón izquierdo del ratón y arrastrando hacia la derecha o izquierda.

Centrados en la parte izquierda, debajo del encabezamiento están los distintos componentes obtenidos con cada procedimiento: log (donde aparece la orden SPSS), título (con el nombre del procedimiento utilizado), notas, archivo de datos utilizado, estadísticos (con el número de casos perdidos y válidos), frecuencias (o el cruce de tablas, o la matriz de correlaciones, etc.) y gráficos solicitados.

Para acceder a los diferentes componentes de la ventana de resultados bastará con hacer clic en uno de estos nombres para que el panel de la derecha seleccione este elemento (mediante un flecha roja en el panel izquierdo o un recuadro en el panel derecho). Este fichero de resultados es fácilmente modificable por el propio SPSS insertando encabezados, títulos, insertando y borrando texto, etc.

Explicadas las partes del visor de resultados, pasaremos a interpretar la distribución de frecuencias de la variable utilizada como ejemplo: la asiduidad con la que los entrevistados utilizan el ordenador (v37). Tras el título del procedimiento (frecuencias) y las notas se muestran los estadísticos solicitados en el cuadro de diálogo de la figura 7.5 (cuartiles, amplitud, mediana y moda). La tabla de resultados *Estadísticos* señala que han sido procesados 191 casos, de los cuales 189 son válidos y 2 perdidos. El valor central de la distribución (mediana) y el valor más frecuente (moda) coinciden en el 1. Más abajo se muestra el rango (valor 98), y los valores que dejan por debajo el 25, 50 y 75% de los casos (cuartiles): se trata del valor 1, el 1 y el 2.

A continuación se presenta otra tabla de resultados con las frecuencias de la distribución. Esta tabla se reproduce en la figura 7.11, y comienza con el nombre y la etiqueta de la variable. Más abajo, en la parte izquierda, aparecen las etiquetas de cada una de las categorías de respuesta tal y como han sido introducidas en el tercer

63. Las versiones del SPSS posteriores a la 14, al permitir utilizar varios archivos de datos, presenta en la primera línea el nombre del archivo de datos con el que se han realizado los análisis.

capítulo. A la derecha la frecuencia absoluta de cada categoría: 97 entrevistados utilizan el ordenador todos o casi todos los días, mientras que 46 lo hacen dos o tres veces a la semana. Si alguna de esas categorías no hubiera sido elegida por ningún entrevistado quedaría eliminada de la tabla.

Pese al interés de la información proporcionada por esta columna, sin duda es mucho más interesante el análisis de la frecuencia relativa o porcentaje al indicar que el 50,8% (97/191) de los entrevistados utilizan el ordenador todos o casi todos los días, mientras que el 24,1% lo hacen dos o tres veces a la semana. La cuarta columna, encabezada con el nombre *porcentaje válido*, se diferencia del *porcentaje* en que no tiene en cuenta los valores perdidos: aquí el *porcentaje válido* tiene en cuenta 189 casos, no considerando dos casos que por diversos motivos han sido eliminados del análisis. En este ejemplo la diferencia entre ambos porcentajes es prácticamente nula, pero cuando los casos perdidos son numerosos las diferencias entre el *porcentaje* y el *porcentaje válido* pueden ser muy elevadas. La interpretación del porcentaje válido es similar a la columna anterior: el 51,3% (97/189) de los que han contestado esta pregunta utilizan el ordenador todos o casi todos los días.

(v37) Frecuencia de utilización: ORDENADOR

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Todos o casi todos los días	97	50,8	51,3	51,3
	Dos o tres veces a la semana	46	24,1	24,3	75,7
	Una vez a la semana	23	12,0	12,2	87,8
	Menos de una vez a la semana	8	4,2	4,2	92,1
	Nunca o casi nunca	12	6,3	6,3	98,4
	No responde	3	1,6	1,6	100,0
	Total	189	99,0	100,0	
Perdidos	Sistema	2	1,0		
Total		191	100,0		

Figura 7.11. Tabla de frecuencias.

En la última columna aparece el porcentaje acumulado, aspecto muy útil en variables ordinales. Así, por ejemplo, si nuestro interés es diferenciar las personas que *habitualmente* utilizan el ordenador[64] el análisis del porcentaje acumulado nos proporciona una rápida visión que el 87,8% de los entrevistados utilizan habitualmente el ordenador, frente al 12,2% que lo hacen con menor frecuencia.

Bajo la tabla de frecuencias está situado el gráfico solicitado (ver ventana izquierda de la figura 7.9), en este caso un gráfico de barras representando los porcentajes (recordar figura 7.6). En la figura 7.12 se aprecia claramente que la mitad de la muestra utiliza el ordenador todos o casi todos los días, y que algo más del 20% lo hace dos o tres veces por semana. Cuando la ventana de resultados contiene algún elemento gráfico, como es este caso, basta con hacer doble clic sobre él para activar la ventana del *Editor de gráficos*.

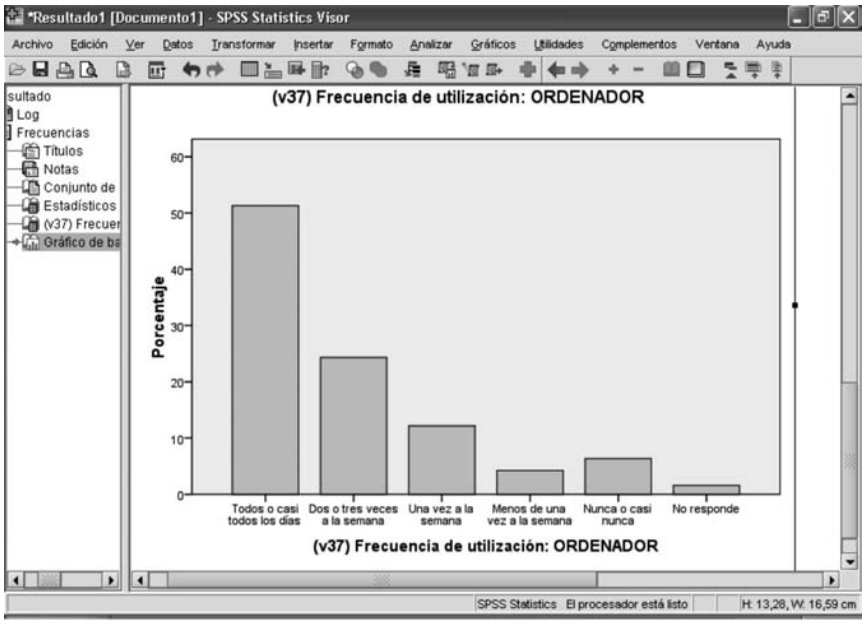


Figura 7.12. Visor de resultados de SPSS.

Terminar señalando que todos los resultados de SPSS pueden insertarse en un editor de texto utilizando el menú *Edición⇒Copiar*. Posteriormente se acude al editor utilizado, Microsoft Word por ejemplo y, situados en el lugar donde se desea copiar, se pro-

64. Considerando "habitualmente" como, al menos, una vez a la semana.

cede con *Edición⇒Pegado especial⇒Imagen*. Cuando lo que se desea es *recuperar* todos los resultados es mejor utilizar la opción *exportar*, que se activa pulsando –dentro del visor de resultados– las opciones el menú *Archivo⇒Exportar*. En la figura 7.13 se muestra el cuadro de diálogo resultante, en el que será necesario cambiar el contenido a exportar (todos, todos visibles o seleccionados), el nombre y el lugar donde se guardará la información exportada, y el formato de exportación. El menú despegable abierto en la parte inferior de la figura 7.13 muestra los formatos disponibles en la versión 17.0.

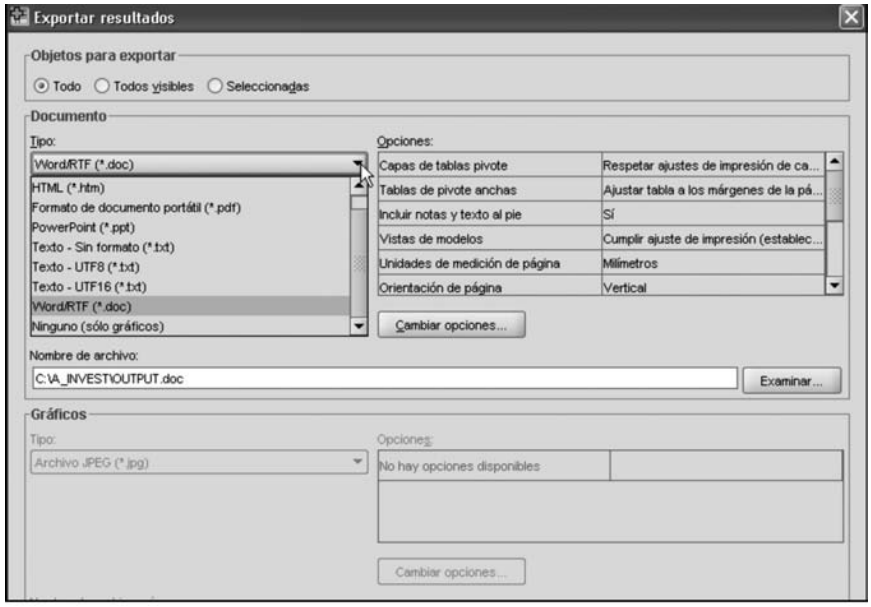


Figura 7.13. Cuadro de diálogo Exportar resultados.

En el *capítulo 7* de los *materiales complementarios* (web) se incluye un estudio sobre *Vida Cotidiana* realizado por la Fundación CIRES hace ya unos años, que será utilizado con el fin de trabajar con una investigación real[65] (totalmente *real*, no como la *simulación de investigación* que venimos utilizando hasta el momento con los cuestionarios respondidos por los estudiantes). Una vez abierto proponemos como ejercicio solicitar las frecuencias de la pregunta 32, variables b55 y b56. Como un segundo ejercicio, para *limar* aspectos que no han podido quedar claros, solicitar las frecuencias de d46, d48 y d50.

65. La utilización de esta investigación nos permite también ver otras formas de codificación.

Antes de finalizar este apartado consideramos importante indicar que se han solicitado las frecuencias de la variable v37 con el fin de *ilustrar* la explicación de la tabla de frecuencias. Téngase en cuenta que el investigador debe comenzar, siempre, analizando la distribución de las *variables básicas de la muestra* con el fin de compararlas con el *universo* del que se ha extraído la información[66]. Estas variables *básicas* aparecen definidas en el diseño muestral, aunque la mayor parte de las veces se refieren al sexo, edad, situación laboral, estado civil, nivel educativo, etc.

Con el fin de *fijar* los conocimientos aprendidos en esta sección –antes de considerar nuevos contenidos– recomendados realizar los ejercicios 1, 2, 3 y 4 incluidos en los *materiales complementarios*, dentro del *capítulo 7*.

4. Análisis de respuestas múltiples categóricas

Pese a que el procedimiento frecuencias es uno de los más utilizados en la investigación con encuesta, en numerosas ocasiones los cuestionarios tienen preguntas que presentan categorías de respuestas no excluyentes, donde los entrevistados pueden seleccionar varias de las alternativas posibles. Son las conocidas como *preguntas multirespuesta* (o *preguntas de respuesta múltiple*), y pueden verse varios ejemplos de éstas en el cuestionario "encuestas estudiantes" mostrado en el apartado 2.6: situaciones que mejor definen la actividad durante el tiempo libre (pregunta 3), asignaturas que proponen libros de lectura obligatoria (pregunta 7), y la pregunta 17a que recoge información sobre los periféricos o dispositivos en el ordenador.

Realizaremos la explicación utilizando como ejemplo la pregunta 3, situaciones que mejor definen la actividad durante el tiempo libre, y cuya información se ha recogido en las variables v03 y v04[67]. El análisis de respuestas múltiples contempla dos procesos: en primer lugar se definen las variables que componen cada pregunta, para posteriormente solicitar las frecuencias o análisis pertinentes.

La definición de las variables que forman la pregunta múltiple se realiza con el menú *Analizar⇒Respuestas múltiples⇒Definir conjuntos* de respuestas múltiples; que da lugar al cuadro de diálogo de la figura 7.14.[68]

66. Como ya señalamos en el segundo párrafo del apartado 2.3 del capítulo II (página 39).
67. Es posible solicitar las frecuencias de cada una por separado y después realizar una suma, si bien no resulta práctico si después se pretende realizar alguna otra operación con esa información.
68. Recuérdese que es posible hacerlo también con el menú *Datos* mostrado en la figura 4.6; concretamente la opción *Definir conjuntos de respuestas múltiples*.

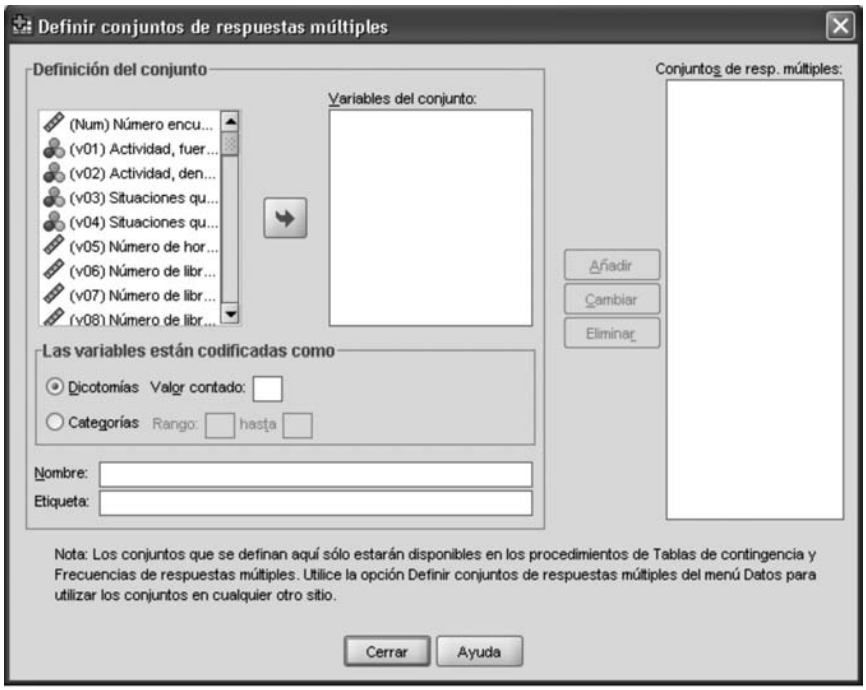


Figura 7.14. Cuadro de diálogo Definir conjuntos de respuestas múltiples.

En la ventana de la izquierda aparecen todas las variables presentes en el archivo de datos, en el centro las variables a definir, y a la derecha las que formarán parte de los conjuntos definidos. Como todavía no hay ninguna, la ventana de la derecha aparece vacía.

Señalamos más atrás que las respuestas a la pregunta 3 se han recogido en las variables v03 y v04, de modo que será necesario pasar ambas variables a la ventana del centro. A continuación se debe indicar la forma de codificación de estas variables, dicotomías o categorías. Las primeras se caracterizan porque presentan dos (o una) respuesta[69], mientras que las segundas se refieren a preguntas con un mayor número de categorías de respuesta.

La pregunta 3 presenta varias opciones de respuesta, de modo que nos encontramos con una pregunta multirespuesta categórica con 11 posibilidades de respuesta (ver libro de códigos del apartado 3.9). Tras marcar *Categorías* en el lugar correspondiente del cuadro de diálogo se anota –en el espacio reservado para este fin– el rango de la variable (menor y máximo valor), en este caso 1 y 11 como vimos en el libro de códigos de la sección 9 del capítulo III (figura 7.15). A continuación debe escribirse un

69. En el siguiente apartado mostraremos un ejemplo de una pregunta multirespuesta dicotómica.

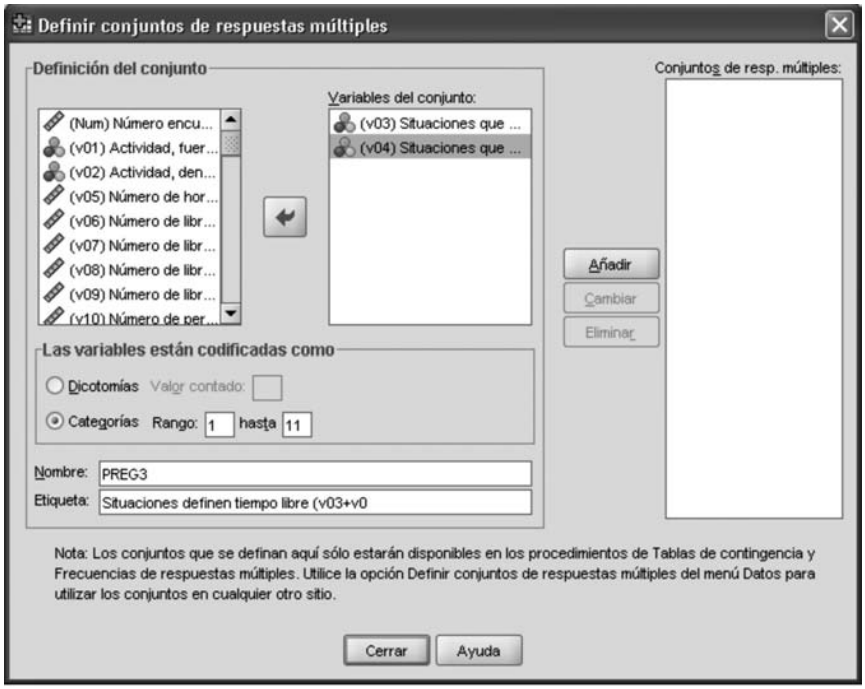


Figura 7.15. Cuadro de diálogo Definir conjuntos de respuestas múltiples, variables categóricas (Pregunta 3).

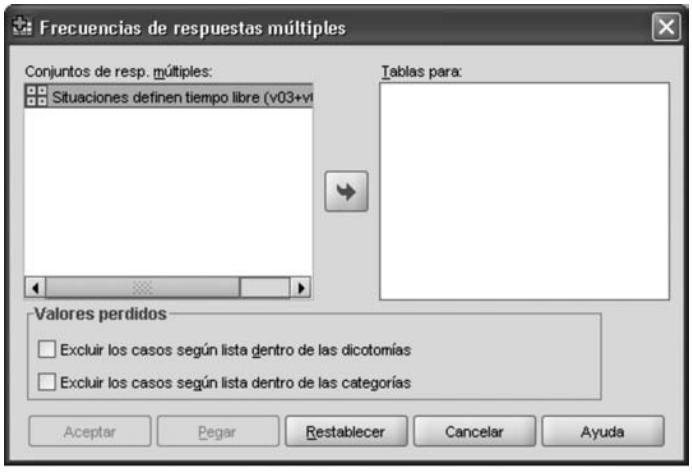


Figura 7.16. Cuadro de diálogo Frecuencias de respuestas múltiples.

nombre para identificar la unión de ambas variables (PREG3 en este caso) y –si se desea– una etiqueta con su descripción, por ejemplo "situaciones que definen el tiempo libre (v03+v04)". Pulsando el botón *Añadir* este nuevo nombre (PREG3) pasará la ventana de la derecha.

Definidos los conjuntos de respuestas múltiples, es el momento de solicitar las frecuencias utilizando para ello el menú *Analizar⇒Respuestas múltiples⇒Frecuencias*. El cuando de diálogo de la figura 7.16 muestra, en su ventana izquierda, los conjuntos de respuestas múltiples definidos, y a la derecha las variables seleccionadas para las frecuencias. Con dos clic de ratón se desplaza la variable recién creada a la ventana de la derecha y, tras pulsar el botón *Aceptar*, aparecen los resultados de la tabla 7.1.

Frecuencias $PREG3

		Respuestas		Porcentaje de casos
		Nº	Porcentaje	
Situaciones definen tiempo libre (v03+v0	Pasarlo bien sin hacer nada	14	3,8%	7,3%
	Hacer muchas cosas	42	11,4%	22,0%
	Dedicarme a las personas más queridas	36	9,8%	18,8%
	Hacer cosas de mi trabajo que tengo pendientes	26	7,0%	13,6%
	Descansar, recuperar fuerzas	41	11,1%	21,5%
	Estar con la gente, charlar, tratar a los amigos	126	34,1%	66,0%
	Aburrirme	6	1,6%	3,1%
	Pensar, meditar	8	2,2%	4,2%
	Dedicarme tranquilamente a mis cosas, aficiones, deportes	70	19,0%	36,6%
Total		369	100,0%	193,2%

Tabla 7.1. Frecuencias de las Pregunta 3.

El nombre de la nueva variable y su etiqueta encabezan los resultados, y más abajo se muestran las etiquetas de la variable, el número de respuestas, el porcentaje respecto al número de respuestas y el número de casos. El porcentaje de respuestas es el ratio

del número de elecciones de cada categoría entre el total de respuestas: 14/369 = 3,8; 42/369 = 11,4; 36/369= 9,8;... El porcentaje de casos, por su parte, corresponde al ratio entre del número de elecciones de cada categoría entre el número de entrevistados, el número de casos válidos: 14/191 = 7,3; 42/191 = 22,0; 36/191= 18,8;...

La interpretación de cada una es distinta, en la medida que la primera centra su atención en las respuestas, y la segunda en los entrevistados. Así del total de respuestas obtenidas *estar con la gente* recibe un 34,1% de respuestas, *dedicarme tranquilamente a mis cosas* un 19%, *hacer muchas cosas* un 11,4%, etc. De todos los entrevistados el 66,0% ocupan su tiempo libre en *estar con la gente*[70], un 36,6% en *dedicarme tranquilamente a mis cosas y aficiones*[71], y otro 22,0% en *hacer muchas cosas*. La experiencia investigadora nos ha demostrado que es mejor utilizar el porcentaje de casos, en la medida que el objeto de estudio son los entrevistados, los sujetos entrevistados, y no el número de respuestas proporcionadas por éstos[72].

Fijar los conocimientos aprendidos es el fin de toda actividad docente. Con el fin de facilitar esta tarea proponemos analizar algunas preguntas de la investigación sobre Vida Cotidiana incluido en los *materiales complementarios*. Concretamente la referida a los objetivos más importantes a solucionar en España (pregunta 6, variables a27, a29 y a31) y en el mundo (pregunta 8, variables a39, a41 y a43)[73]. Nuestro interés se centra también en conocer los tipos de personas –al margen de los familiares– con los que el entrevistado se relaciona habitualmente; pregunta 34, variables b63, b64 y b65.

5. Análisis de respuestas múltiples dicotómicas

En la figura 7.14 y en la explicación de las respuestas múltiples categóricas, dimos cuenta de la existencia de otro tipo de respuestas múltiples conocidas como dicotómicas. Se trata, por ejemplo, de las respuestas de la pregunta 14a (variables v22–V29) donde cada entrevistado es consultado por los dispositivos presentes en el ordenador, y su respuesta es *si* (codificada con el valor 1) o *no* (codificada con el valor 0). Algún lector estará pensando que es posible analizarla solicitando tablas de frecuencias para

70. El 34% (100–66,0) restante ha elegido otras "situaciones" que definen su tiempo libre.
71. El 63,4% (100–36,6) restante ha elegido otras "situaciones" que definen su tiempo libre.
72. Recomendamos leer un artículo de prensa que se encuentra en los *materiales complementarios* para observar cómo se ha presentado una pregunta similar en un medio de comunicación.
73. Considerando que se trata de tres *objetivos* igualmente importantes, no ordenados según la mayor o menor trascendencia. Esto es, olvidando que la variable a39 recoge el objetivo más importante, la a41 el segundo más importante, y la a43 el tercero más importante.

cada una de las variables que componen esta pregunta, y la verdad es que no estará del todo desacertado, pero existe una forma mejor de hacerlo utilizando la *definición de respuestas múltiples dicotómicas*.

La definición de los conjuntos de respuestas múltiples se lleva a cabo de la misma forma que vimos en el apartado anterior (*Analizar⇒Respuestas múltiples⇒Definir conjuntos...*), dando lugar al cuadro de diálogo de la figura 7.14. Ahora bien, a diferencia del procedimiento expuesto en el apartado anterior, en este caso se procede a definir las variables como *Dicotomías*, indicando el valor que se desea contar. Como la posesión de dispositivos en el ordenador se ha codificado con el valor 1, procederemos a contar este valor (ver figura 7.17).

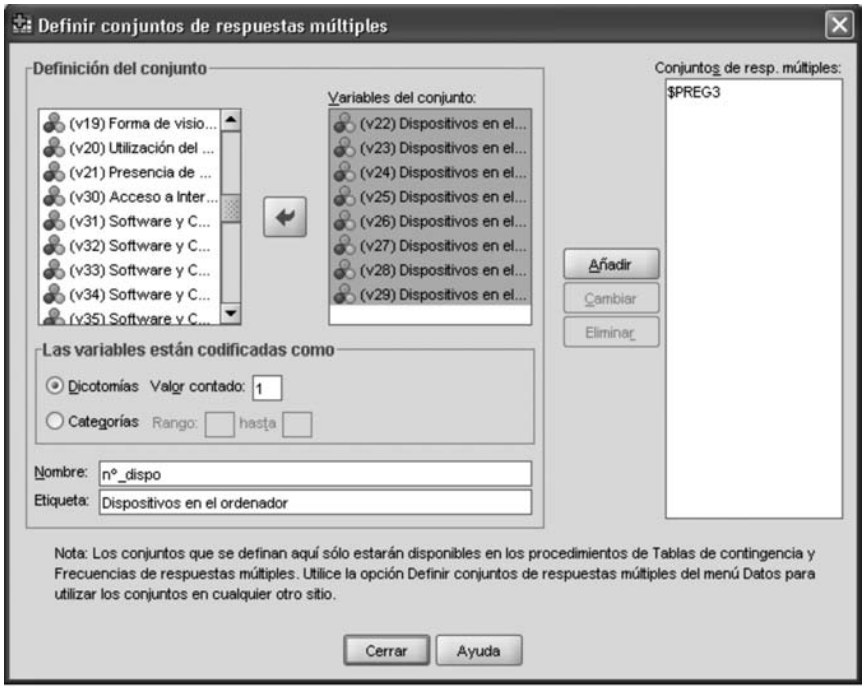


Figura 7.17. Cuadro de diálogo Definir conjuntos de respuestas múltiples, variables dicotómicas (Pregunta 17a).

A continuación se elige un nombre para la variable resultante (por ejemplo Nº_dispo), y una etiqueta que ayude a su identificación. Pulsando el botón *Añadir* este nuevo nombre pasará al recuadro de la derecha, que como vemos ya no se encuentra vacío sino que está presente la variable definida en el apartado anterior.

Definido el conjunto de respuestas múltiples dicotómicas procederemos a solicitar las frecuencias utilizando el menú *Analizar⇒Respuestas múltiples⇒Frecuencias*. El cuando de diálogo de la figura 7.18 muestra, en su ventana izquierda, los conjuntos de respuestas múltiples definidos, y a la derecha las variables seleccionadas para las frecuencias. Con dos clic de ratón se desplaza la variable Nº_dispo a la ventana de la derecha y, tras pulsar el botón *Aceptar*, aparecerán los resultados de la tabla 7.2.

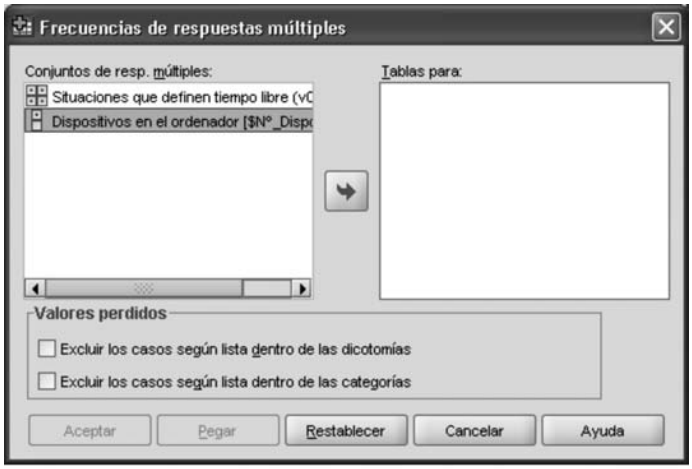


Figura 7.18. Cuadro de diálogo Frecuencias de respuestas múltiples.

No entraremos en una interpretación detallada de los resultados puesto que en la tabla 7.1 ya se ha señalado la diferencia entre el porcentaje de respuestas y de casos. Tan sólo indicar que de los entrevistados con ordenador el 94,4% dispone también de impresora, ausente en el 5,6% (100 – 94,4) de los equipos. El 92,2% cuenta con lectora de CD (ausente en el 7,8% de los ordenadores) y el 91,1% con altavoces (no presente en el 8,9%). Ahora bien, de los *equipamientos* presentes en los ordenadores de los entrevistados, el 19,6% son impresoras, el 19,2% lectoras de CD, un 18,9% altavoces, etc. Dicho de otro modo, utilizar el *porcentaje de respuestas* implica describir la distribución del número de equipamientos totales, los 866 equipamientos declarados por los 191 entrevistados.

Terminar esta sección destacando una de las limitaciones que presentan estas preguntas; como es la necesidad de definir los conjuntos de respuestas múltiples cada vez que comienza una nueva sesión de trabajo. Dicho de otro modo, los conjuntos de respuestas múltiples no pueden ser grabados en el archivo de datos cuando se ter-

Frecuencias $nº_dispo				
		Respuestas		Porcentaje de casos
		Nº	Porcentaje	
Dispositivos en el ordenador	(v22) Dispositivos en el ordenador: IMPRESORA	170	19,6%	94,4%
	(v23) Dispositivos en el ordenador: MODEM	134	15,5%	74,4%
	(v24) Dispositivos en el ordenador: ALTAVOCES	164	18,9%	91,1%
	(v25) Dispositivos en el ordenador: WEBCAM	36	4,2%	20,0%
	(v26) Dispositivos en el ordenador: LECTORA CD	166	19,2%	92,2%
	(v27) Dispositivos en el ordenador: GRABADORA CD	68	7,9%	37,8%
	(v28) Dispositivos en el ordenador: LECTORA DVD	104	12,0%	57,8%
	(v29) Dispositivos en el ordenador: GRABADORA DVD	24	2,8%	13,3%
Total		866	100,0%	481,1%

Tabla 7.2. Frecuencias de las Pregunta 14a.

mina la sesión[74]. Esta restricción, junto con una interpretación más complicada de los resultados, la ausencia de gráficos en los cuadros de diálogo, y la imposibilidad de utilizar estadísticos para conocer la relación entre variables[75] recomienda utilizar con mesura este tipo de preguntas[76].

Buscando *fijar* los conocimientos aprendidos en esta sección, antes de considerar nuevos contenidos, proponemos analizar dos preguntas de la investigación sobre Vida Cotidiana. Concretamente el equipamiento disponible en los hogares, que se reco-

74. Sin embargo, es posible definir varios conjuntos abriendo una sola vez el cuadro de diálogo *Definir conjuntos de respuestas múltiples*. Para ello, tras definir la primera variable (por ejemplo PREG3) y pulsar el recuadro *Añadir*, se procede con la definición de la segunda sin necesidad de cerrar este cuadro de diálogo. Digamos que tras pulsar *Añadir* se pone fin a la definición de cada conjunto de respuestas múltiples.
75. La relación entre variables se analizará en el capítulo noveno.
76. Otras limitaciones se presentarán en el décimo capítulo, dentro de apartado 2 del capítulo X.

ge en la pregunta 20 (variables de la b8 a la b20), con el valor 1 (esto implica, lógicamente, considerar únicamente este valor). Otro aspecto que nos interesa es conocer la frecuencia con la que se comen determinados productos alimenticios (pregunta 56, variables de la c58 a la c64), considerando únicamente aquellos entrevistados que "comen todos o casi todos los días" (opción 1).

6. Estadísticos descriptivos

Este capítulo sobre la obtención de información finaliza presentando los estadísticos utilizados para variables numéricas (escalas de intervalo), utilizando el procedimiento estadísticos descriptivos[77]. Es preciso señalar que se trata de un tipo de variables no muy frecuentes en la investigación mediante encuesta.

Marcando *Analizar⇒Estadísticos descriptivos⇒Descriptivos* aparece el cuadro de diálogo de la figura 7.19, utilizado para obtener información de variables numéricas o escalas de intervalo. Hasta ahora siempre se ha trabajado con las variables de una a una, si bien es posible solicitar varias en cada procedimiento. De hecho, en la figura 7.19 se solicitan los estadísticos descriptivos de V05 (número de horas libres que dispone a la semana para ocio o diversión) y v06 (número de libros leídos relacionados con los estudios).

Pulsando el botón *Opciones*, situado en la parte inferior derecha de la figura 7.19, aparece el cuadro de diálogo secundario para seleccionar los estadísticos pertinentes. En este caso se dejan los estadísticos *por defecto*; la media, desviación típica, valor máximo, mínimo y la suma. Pulsando consecutivamente los botones *Continuar* (cuadro de diálogo secundario) y *Aceptar* (cuadro de diálogo *Descriptivos*) aparecerá, en el menú de resultados, la figura 7.21.

El número de horas libres a la semana en la muestra seleccionada oscila entre 7 y 80 (figura 7.21), presentando un valor medio de 38,44 con una variabilidad de 20,179. Las personas entrevistadas han tenido –en total– 7.149 horas libres. Es reseñable que cinco personas no han contestado a esta pregunta, puesto que de 191 personas entrevistadas se han recibido 186 respuestas. Por otro lado, el número de libros leídos relacionados con los estudios fluctúa entre 0 y 15, presentando un valor medio de 2,69 y una desviación típica de 2,883. Las 191 personas entrevistadas han leído –en total– 514 libros.

77. Es posible analizar los estadísticos para variables numéricas utilizando el procedimiento *Explorar*, que no trataremos aquí por haberlo desarrollado en un trabajo anterior (Díaz de Rada, 2002).

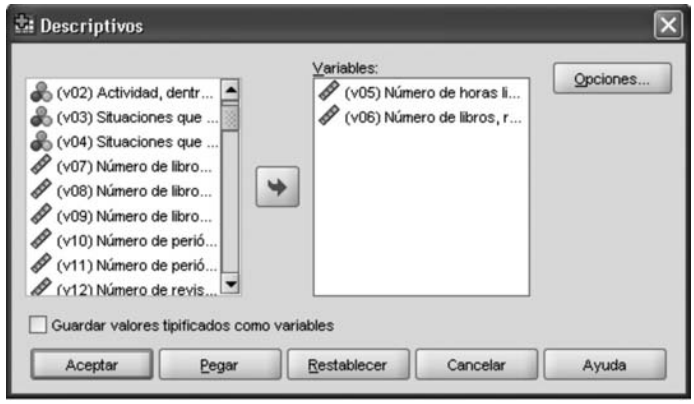


Figura 7.19. Cuadro de diálogo Descriptivos.

Figura 7.20. Cuadro de diálogo Descriptivos: opciones.

Ahora bien, si se eliminan las personas que no leen libros (valor 0 en v06), el tamaño muestral desciende de 191 a 152 personas; lo que implica que 39 entrevistados (respecto a 191 supone un 20,7% de la muestra) no ha leído ningún libro en el período observado. Esto implica, considerando los 514 libros leídos, una media de 3,38

	N	Mínimo	Máximo	Suma	Media	Desv. típ.
(v05) Número de horas libres que dispone a la semana de ocio o diversión	186	7	80	7149	38,44	20,179
(v06) Libros relacionados con tus estudios leídos en el último año	191	0	15	514	2,69	2,883
N válido (según lista)	186					

Figura 7.21. Estadísticos descriptivos de la variable v05 y v06: Número de horas libres que dispone a la semana para ocio o diversión y libros relacionados con tus estudios leídos en el último año.

	N	Mínimo	Máximo	Suma	Media	Desv. típ.
(v06) Libros relacionados con tus estudios leídos en el último año	152	1	15	514	3,38	2,847
N válido (según lista)	152					

Figura 7.22. Estadísticos descriptivos de la variable v06: Libros relacionados con tus estudios leídos en el último año.

libros por entrevistado que lee libros, y una variabilidad de 2,847, como se muestra en la figura 7.22.

Con el fin de fijar los conocimientos aprendidos en esta sección, antes de considerar nuevos contenidos, recomendados interpretar la "evaluación de la situación personal *actual* del entrevistado, retrospectiva y prospectiva a un año" (pregunta 5 de la investigación sobre Vida Cotidiana, variables a21, a23 y a25). La valoración personal, ¿es mejor o peor que la valoración sobre la situación del país? (pregunta 7, variables a33, a35 y a37). ¿Y respecto al mundo? (pregunta 9).

7. Anexo 1: Lenguaje de sintaxis de SPSS con los análisis realizados en el capítulo

Los procedimientos presentados en este apartado se han elaborado pulsando el botón *Pegar* del cuadro de diálogo *Frecuencias* presentado en la figura 7.1. El texto con las órdenes escritas utilizando el lenguaje de sintaxis[78] puede ser guardado y recuperarse en sesiones posteriores. La ejecución de estas órdenes se realizan, desde la ventana de sintaxis, seleccionando el menú *Ejecutar* y eligiendo una de las posibilidades disponibles: *Todo*, *Selección*, *Actual* (Ctrl+R) ó *Hasta el final*. Otra forma más rápida y sencilla es, situados en una determinada orden, pulsar el incono .

Apartado 2: Frecuencias de variables nominales y ordinales (figura 7.8).

```
FREQUENCIES
VARIABLES=v37
/NTILES= 4
/STATISTICS= RANGE MEDIAN MODE
/BARTCHART PERCENT
/ORDER= ANALYSIS.
```

Apartado 4: Análisis de respuestas múltiples categóricas

```
MULT RESPONSE
GROUPS=$PREG3 'Situaciones definen tiempo libre (v3 +v4) (v03 v04 (1,11))
/FREQUENCIES=$PREG3.
```

Apartado 5: Análisis de respuestas múltiples dicotómicas

```
MULT RESPONSE
GROUPS=$Nº_dispo 'Dispositivos en el ordenador' (v22 v23 v24 v25 v26 v27
    v28 v29 v30 (1))
/FREQUENCIES=$Nº_dispo.
```

78. Otros lo denominan "lenguaje de comandos".

Apartado 6: Estadísticos descriptivos

```
DESCRIPTIVES
VARIABLES=v06
/STATISTICS=MEAN SUM STDDEV MIN MAX.
```

Capítulo VIII

Transformación de datos y creación de nuevas variables

1. Objetivos didácticos del capítulo

Dedicaremos este capítulo a presentar una serie de procedimientos utilizados para realizar cambios en los valores de las variables, llevando a cabo transformaciones de la distribución de frecuencias original[79]. Las diferencias entre todos ellos se fundamentan en el criterio de transformación, si es decidido por el programa SPSS (*recodificación* automática y *categorizar variables*) o por el investigador. Cuando es el usuario el que decide el criterio de transformación es posible diferenciar entre unión-agregación de categorías (recodificar), realizar cálculos (calcular), y creación de nuevas variables uniendo determinados valores de otras variables (contar valores dentro los casos). En todos éstos al terminar la sesión de trabajo el SPSS preguntará si desea guardar el contenido de la ventana de datos, puesto que se han producido cambios en el archivo de datos. Será preciso guardarlo si se desean utilizar estas transformaciones en otro momento. De no hacerlo se perderán todos los cambios realizados durante esta sesión de trabajo.

Un apartado dedicado a la selección de casos mediante criterios condicionales pondrá fin a este capítulo. Recuérdese las razones que justifican la creación de nuevas variables, presentadas en el último párrafo de la sección 2.3. del capítulo II.

Como en anteriores capítulos la explicación se llevará a cabo utilizando el archivo de datos obtenido del cuestionario presentado en el segundo capítulo, sección 2.9 (ENCUESTAS ESTUDIANTES 2002_03.SAV). Los ejercicios propuestos en el capítulo ocho de los *materiales complementarios* (web) deberán realizarse con el archivo "Encuestas estudiantes (SIETE promociones).sav".

79. Recuérdese que en la sección 2.3 del capítulo II (página 39) se presentaron las razones que justifican estos procesos.

2. Recodificación automática

Comenzaremos con la *recodificación* automática, procedimiento que consiste en crear una nueva variable donde los valores numéricos y de cadena (de una variable original) se cambian a valores enteros consecutivos. Es una forma sencilla de convertir las variables de cadena en variables numéricas conservando los atributos de las primeras (etiquetas, ancho de columna, etc.), y resulta especialmente apropiado para los procedimientos que no pueden utilizar variables de cadena o que precisan valores enteros consecutivos. Determinados procedimientos estadísticos necesitan valores secuenciales, puesto que las casillas vacías reducen el rendimiento, al tiempo que precisan de más memoria del ordenador.

Este procedimiento se encuentra dentro del menú Transformar, de modo que para llevarlo a cabo será preciso seleccionar *Transformar⇒Recodificación automática*, tras lo cual aparecerá el cuadro de diálogo de la figura 8.1. Tras seleccionar una variable (en este caso V41) se escribe –en el espacio correspondiente– el nombre que tendrá la nueva variable categorizada. Tras pulsar el botón *Nuevo nombre* éste sube a la ventana superior, al tiempo que los recuadros *Aceptar* y *Pegar* cambian de color, lo que indica que ya se puede proceder con la *recodificación* (figura 8.2). Es posible realizar la recodficación comenzando con el menor valor o por el valor superior.

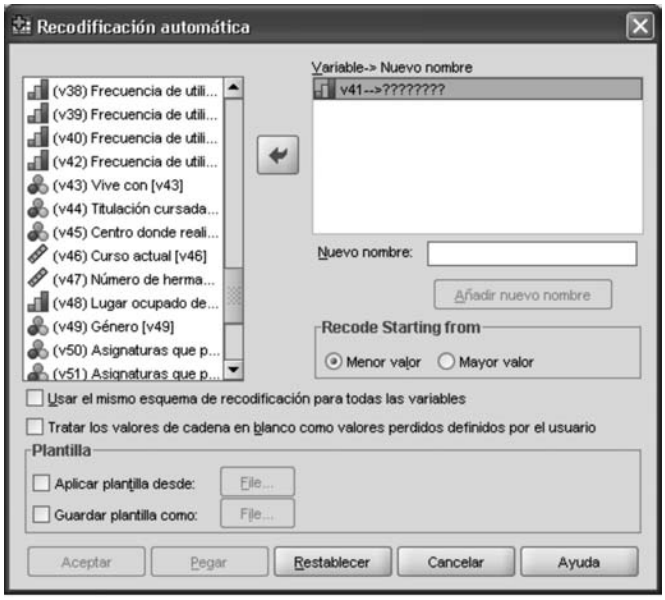


Figura 8.1. Cuadro de diálogo Recodificación Automática.

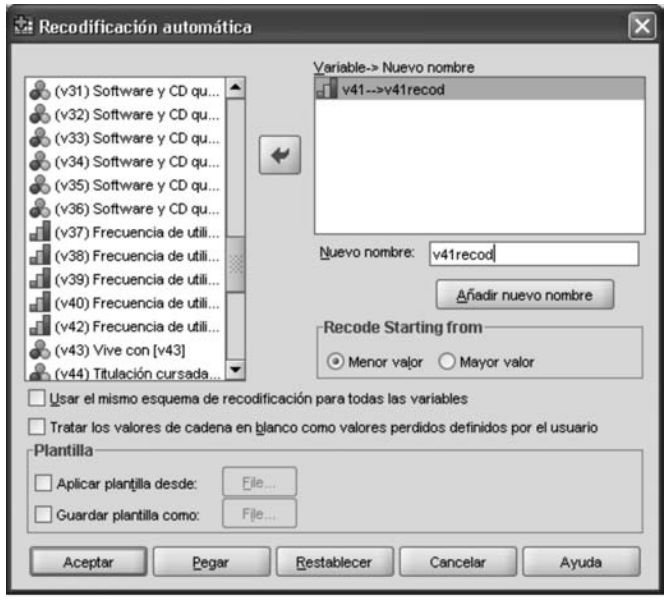


Figura 8.2. Cuadro de diálogo Recodificación automática con la variable v41recod.

Tras pulsar el botón *Aceptar* se crea una nueva variable en la ventana de datos, al tiempo que aparece una tabla (en el visor de resultados) donde se muestran los valores antiguos, los nuevos y las etiquetas de valor. En esta tabla, que se muestra en la tabla 8.1, se aprecia que ninguna de las personas entrevistadas utiliza todos los días una hoja de cálculo, es decir que nadie había elegido la opción 1 en esta variable. Esta situación explica por qué procedemos con la recodificación automática de esta variable[80].

En la tabla 8.1 se aprecia cómo el valor 2 se ha convertido en el 1, el 3 en 2, el 4 en 3, el 5 en 4, y el 99 en 5. De modo que será necesario codificar este valor como valor perdido antes de proceder con el análisis de esta nueva variable. Por último, debe introducirse esta nueva variable en el libro de códigos (ver documento *libro de códigos final* en los *materiales complementarios*); recomendando situarla cerca de la variable original, esto es, entre v41 y v42.

80. No estará de más recordar que, en la sección 4 del capítulo III dedicada a la explicación de los tipos de variables considerando la escala de medida, se recomendaba que la medición ordinal debe respetar las relaciones observadas en la asignación del sistema de medición, ordenando los números según su orden serial.

V41 Old Value	V41RECOD New Value	Frecuencia de utilización: HOJA DE CALCULO Value Label
2	1	Dos o tres veces a la semana
3	2	Una vez a la semana
4	3	Menos de una vez a la semana
5	4	Nunca o casi nunca
99	5	No responde

Tabla 8.1. Resultado de la Recodificación automática.

Este programa siempre sitúa las *nuevas variables* al final del archivo de datos, a la derecha de la última variable. Recomendamos colocar las nuevas variables *cerca* de la variable original, con el fin de tener clara su procedencia. Actuar de este modo evitará confusiones en análisis posteriores.

Para ello, una vez creada la nueva variable, ésta se selecciona y –utilizando el menú *Edición⇒Cortar*– se elimina del final del archivo de datos[81]. Posteriormente debe crearse un *nuevo espacio* (variable) para colocar estos nuevos datos, optando por insertar una nueva variable a la derecha de v41 (por seguir con el ejemplo utilizado). Para ello basta con seleccionar V42 y pulsar el menú *Datos⇒Insertar variable*. Más sencillo resulta, una vez seleccionada la variable v42, pulsar –dentro de la barra de herramientas– el icono . Una segunda forma de insertar la variable es, una vez seleccionada la variable v41, pulsar el botón secundario (derecho) del ratón y marcar la opción "insertar variable. Se opte por cualquiera de estos procedimientos, aparecerá una nueva variable entre v41 y v42, con el nombre var00002. Seleccionada esta variable, bastará con ejecutar el menú *Edición⇒Pegar* para colocar aquí la variable recién creada (en este caso, v41recod).

Fijar los conocimientos aprendidos es el fin de toda actividad docente. Con el fin de facilitar esta tarea proponemos la realización de algunas transformaciones en los datos de la investigación sobre *Vida Cotidiana* que se encuentra en los *materiales complementarios*. Concretamente proponemos realizar una *recodificación automática* de la pregunta que solicita información sobre el número de *amigos de verdad* que tienen los entrevistados (pregunta 34b, variable B68). Realizar esta misma operación en la pregunta que muestra el número de horas que trabajan los entrevistados (pregunta 39, b74).

81. Recuérdese que, al igual que el resto de aplicaciones que funcionan bajo entrono Windows, el menú *Edición⇒Cortar* hace que esta variable desaparezca del archivo de datos, para permanecer guardada en la memoria del ordenador.

3. Agrupación visual

La categorización de variables consiste en transformar variables cuantitativas en cualitativas, creando una nueva variable que contenga la información categorizada de la primera[82]. Categorizar es, en definitiva, convertir datos *numéricos continuos* en un *número discreto* de categorías, convertir variables numéricas (intervalo) en ordinales agrupando dos o más valores contiguos en una misma categoría. Este procedimiento se utiliza también para reducir el número de categorías de las variables ordinales.

Explicaremos el funcionamiento de este procedimiento con un ejemplo, concretamente la variable v05 donde se recoge el número de horas libres a la semana para ocio o diversión. Deseamos dividirla en cuatro categorías. Antes de llevar a cabo esta operación es preciso conocer su distribución. En la tabla 8.2 se muestra la distribución de esta variable, obtenida tras solicitar sus frecuencias (*Analizar⇒Estadísticos descriptivos ⇒Frecuencias*).

Para categorizar esta variable se abre el cuadro de diálogo que se muestra en la figura 8.3 mediante el menú *Transformar⇒Categorizador Visual* y, tras seleccionar la variable, se arrastra a la ventana de la derecha (ventana Categorizar variables) o se marca el *botón-flecha* entre ventanas. Tras pulsar el botón *Continuar* aparece el cuadro de diálogo principal de la categorización (figura 8.4). Cuando se selecciona con el ratón la variable a categorizar ésta pasa automáticamente a ocupar las ventanas superiores del cuadro de diálogo (ver figura 8.5), al tiempo que se muestra su distribución mediante un histograma parte central de la figura 8.5. En este momento deberá escribirse el nombre la nueva variable, por ejemplo v05_ca_1.

La creación de las categorías se realiza pulsando el botón *Crear puntos de corte*, que da acceso al cuadro de diálogo de la figura 8.6. La categorización puede realizarse de dos formas:

- La primera, que es la que aparece seleccionada por defecto, consiste en crear intervalos de igual amplitud. Para ello debe introducirse el primer punto de corte, el número de puntos de corte, la amplitud de cada categoría, y la posición del último punto de corte. En este ejemplo, que recordemos buscaba categorizar v05 en cuatro categorías de igual amplitud, debemos considerar que se trata de una variable con un *recorrido* entre el valor 7 y el 80[83], como hemos podido apre-

82. Volveremos a insistir en algo ya señalado en el apartado anterior: determinados procedimientos precisan valores secuenciales, puesto que las casillas vacías reducen el rendimiento y ocupan más memoria del ordenador.

83. Es necesario que previamente se haya definido el valor 999 (no responde) como valor perdido, puesto que de lo contrario cambiará la amplitud de cada categoría y la posición del último punto de corte.

(v05) Número de horas libres que dispone a la semana de ocio o diversión					
		Frecuencia	**Porcentaje**	**Porcentaje válido**	**Porcentaje acumulado**
Válidos	7	2	1,0	1,1	1,1
	8	2	1,0	1,1	2,2
	10	10	5,2	5,4	7,5
	12	2	1,0	1,1	8,6
	14	4	2,1	2,2	10,8
	15	3	1,6	1,6	12,4
	16	3	1,6	1,6	14,0
	18	2	1,0	1,1	15,1
	20	26	13,6	14,0	29,0
	24	4	2,1	2,2	31,2
	25	12	6,3	6,5	37,6
	30	10	5,2	5,4	43,0
	35	8	4,2	4,3	47,3
	36	2	1,0	1,1	48,4
	40	26	13,6	14,0	62,4
	43	2	1,0	1,1	63,4
	44	4	2,1	2,2	65,6
	45	4	2,1	2,2	67,7
	48	6	3,1	3,2	71,0
	50	12	6,3	6,5	77,4
	55	4	2,1	2,2	79,6
	58	2	1,0	1,1	80,6
	60	4	2,1	2,2	82,8
	61	2	1,0	1,1	83,9
	62	2	1,0	1,1	84,9
	63	2	1,0	1,1	86,0
	68	2	1,0	1,1	87,1
	70	6	3,1	3,2	90,3
	72	8	4,2	4,3	94,6
	78	4	2,1	2,2	96,8
	80	6	3,1	3,2	100,0
	Total	186	97,4	100,0	
Perdidos	No responde	5	2,6		
Total		191	100,0		

Tabla 8.2. Tabla de frecuencias de v05.

Figura 8.3. Categorizar variables.

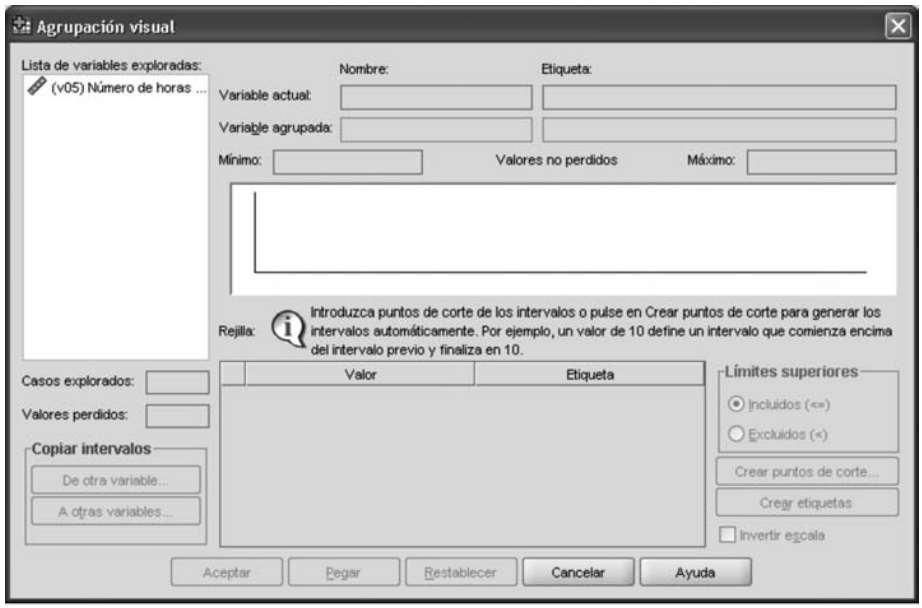


Figura 8.4. Cuadro de diálogo principal de *Categorizar* variables.

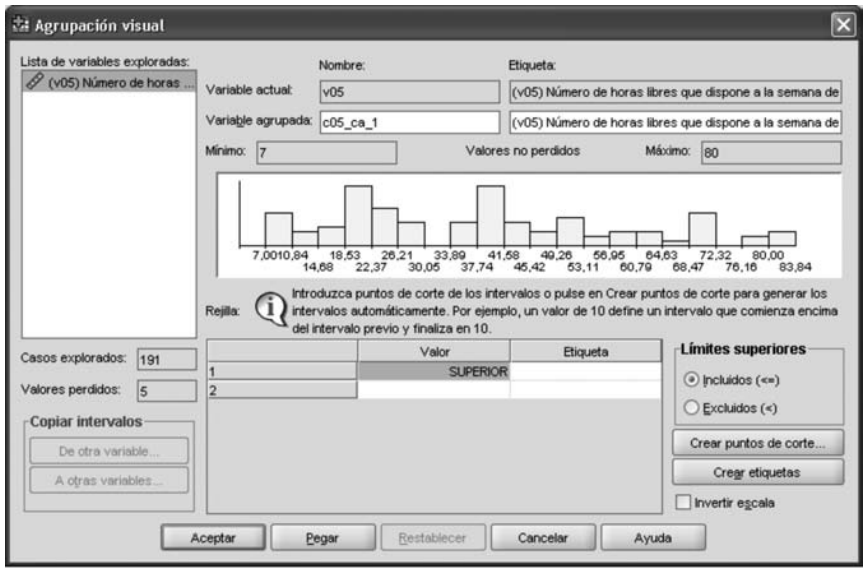


Figura 8.5. Categorizar variables: distribución de la variable a categorizar.

ciar en la parte izquierda de la tabla 8.2. Esta información se muestra también en el cuadro de diálogo principal de la categorización visual, justo encima del histograma (ver figura 8.5).

Observando la *parte izquierda* de la tabla 8.2, dividir en dos partes esta distribución cuyo valor máximo es el 80 implicaría *cortar* la distribución por el punto 40. Para dividirla en cuatro partes –que es nuestro objetivo– deberíamos considerar como puntos de corte el 20, el 40 y el 60[84]; de modo que los intervalos serán 0–20, 21–40, 41–60, y más de 60. El primer punto de corte, según esta argumentación, sería el 20, distribuyendo el resto en cuatro grupos de 20 casos ([(80-20)/3] = 20); cifra que indica la amplitud de cada categoría.

Para llevar a cabo esta categorización debe indicarse al programa que se desean hacer tres puntos de corte y, posteriormente, puede optarse por colocar la anchura (20) o el primer punto de corte (20). El programa precisa de dos informaciones y el calcula el resto: así, tras introducir el primer punto de corte (20) y el número de puntos de corte (3), bastará con pulsar el tabulador para que el programa calcule la amplitud y la *posición del último punto de corte* (figura 8.6)[85].

84. Téngase en cuenta que el número de puntos de corte es el número de categorías menos uno.
85. Cuando se introduce el número de puntos de corte y la amplitud, el programa calcula el primer y el último punto de corte.

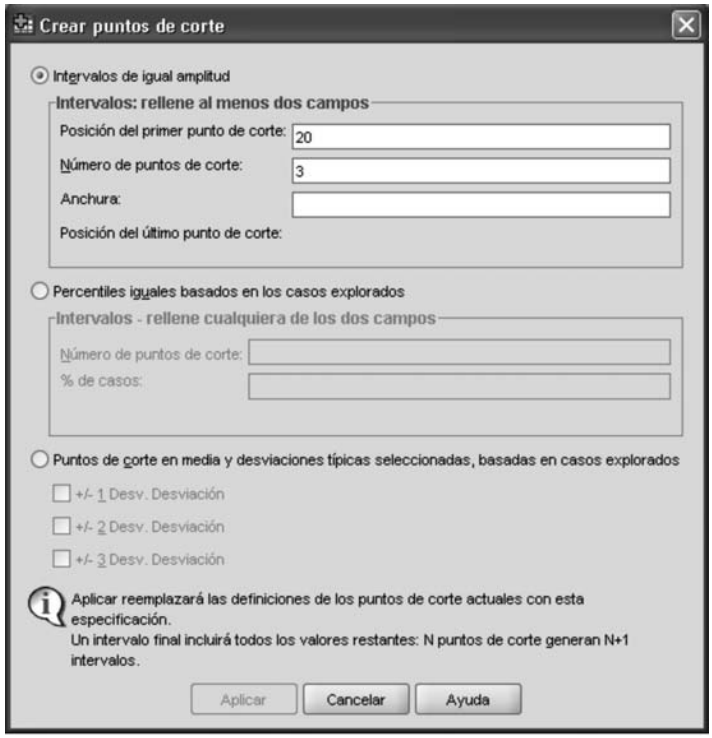


Figura 8.6. Categorizar variables: crear puntos de corte.

Tras introducir esta información el botón *Aplicar* cambia de color para indicar que ya es posible proceder.

Una vez pulsado este botón se vuelve al cuadro de diálogo de la figura 8.7 donde el programa presenta los puntos de corte (límites superiores de cada intervalo), representados gráficamente en el histograma con tres líneas verticales. Se procede de esta forma con el fin de realizar una primera valoración, permitiendo volver a redefinir los puntos de corte cambiando los valores, o bien arrastrando con el ratón las líneas del histograma.

- La segunda forma realiza una categorización considerando el número de casos. Situados en el centro del cuadro de diálogo de la figura 8.6, concretamente en *percentiles iguales basados en casos explorados*, basta con introducir el número de puntos de corte, 3 en nuestro ejemplo, y el programa calculará automáticamente el porcentaje de casos (25% en este ejemplo). Pulsando el botón *Aplicar* se podrá apreciar la división en el histograma del cuadro de diálogo principal (figura 8.8).

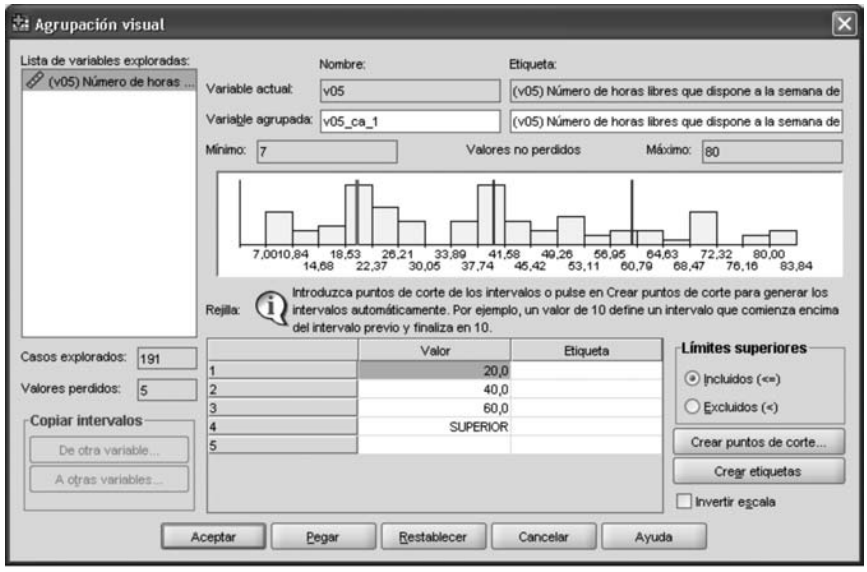


Figura 8.7. Categorizar variables, crear intervalos igual longitud: histograma con tres puntos de corte (cuatro categorías) de igual amplitud.

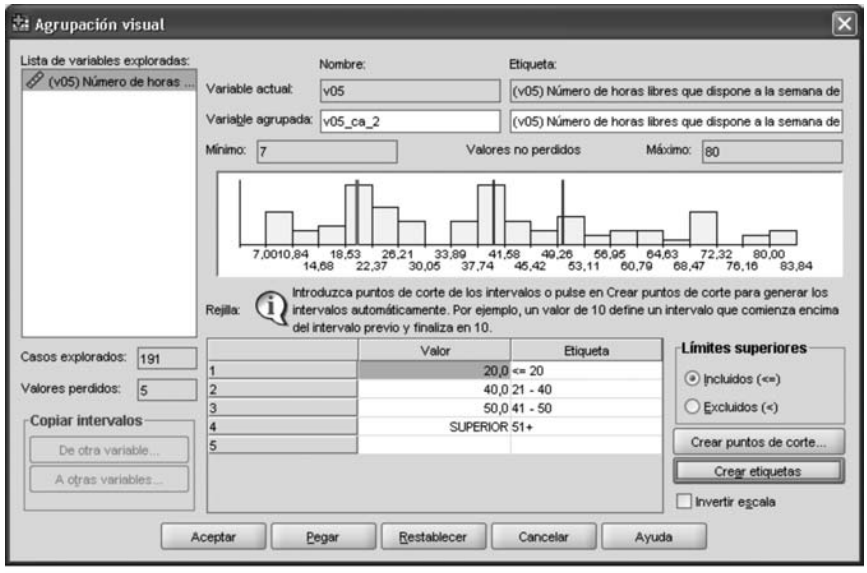


Figura 8.8. Categorizar variables, percentiles iguales: crear etiquetas.

Elegida una determinada categorización recomendamos pulsar el botón *Crear etiquetas*, situada debajo del botón *Crear puntos de corte*, para que el programa añada las etiquetas correspondientes a cada una de las variables categorizadas (ver figura 8.8). Debajo de este botón el programa permite invertir la escala; que consiste en que los valores superiores se conviertan en inferiores y estos últimos en superiores. Por último, pulsando el botón *Aceptar* el programa que creará una nueva variable en la ventana de datos.

(v05) Número de horas libres que dispone a la semana de ocio o diversión (Categorizada)

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	<= 20	54	28,3	29,0	29,0
	21 - 40	62	32,5	33,3	62,4
	41 - 60	38	19,9	20,4	82,8
	61+	32	16,8	17,2	100,0
	Total	186	97,4	100,0	
Perdidos	No responde	5	2,6		
Total		191	100,0		

Tabla 8.3a. Resultado de Categorizar Variables: crear intervalos de igual amplitud (amplitud 20).

(v05) Número de horas libres que dispone a la semana de ocio o diversión (Categorizada)

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	<= 20	54	28,3	29,0	29,0
	21 - 40	62	32,5	33,3	62,4
	41 - 50	28	14,7	15,1	77,4
	51+	42	22,0	22,6	100,0
	Total	186	97,4	100,0	
Perdidos	No responde	5	2,6		
Total		191	100,0		

Tabla 8.3b. Resultado de Categorizar Variables: percentiles iguales basados en los casos explorados.

En este texto se han llevado a cabo las dos categorizaciones explicadas, denominando a la primera *nueva* variable con el nombre v05_ca_1 y a la segunda con v05_ca_2. Solicitadas las frecuencias, que se presentan en las tablas 8.3a y 8.3b, pueden apreciarse las semejanzas y diferencias entre ambas. La tabla 8.3a muestra la variable v05 dividida en cuatro categorías con *intervalos de igual amplitud* (resultado del cuadro de diálogo de la figura 8.6), una amplitud de 20 casos en cada categoría. En la tabla 8.3b se presenta la variable v05 categorizada mediante *percentiles iguales basados en los casos explorados*. Recuérdese que el objetivo era dividir la variable en cuatro categorías con el 25% de los casos en cada una, pero el elevado número de personas que disponen de 20 horas libres a la semana (un 14% según se aprecia en la tabla 8.2) impide realizar el primer corte en el percentil 25, y lo hace en el 29. Algo parecido sucede en el segundo cuartil (percentil 50), puesto que 26 entrevistados (de 186) manifiestan tener 40 horas libres.

Sintetizando, la primera categorización (*intervalos de igual amplitud*) se lleva a cabo considerando la "parte izquierda" de la distribución (donde se muestra la *longitud* de los intervalos, haciendo *cortes* en 20, 40 y 60), mientras que en la segunda tiene en cuenta la "parte derecha" de la distribución, el porcentaje de respuestas (cortando cada 25% de los casos); como se muestra en la tabla 8.4.

La ventaja de esta última (mediante *percentiles iguales basados en los casos explorados*) es la similitud del número de respuestas en cada categoría; esto es, que impide la existencia de categorías con pocos casos[86]. Esto implica que esta herramienta es muy adecuada cuando se analizan variables asimétricas; aquellas que presentan muchos valores en las primeras categorías de la variable y pocos en las últimas (o al revés, pocos en las primeras y muchos en las últimas).

Así sucede –por ejemplo– con el "número de libros relacionados con tus estudios leídos en el último año" (v06), cuya distribución se presenta en la tabla 8.5: un 20,4% de los entrevistados responde ninguno, un 25,1% lee un libro, un 14,7% dos, un 13,6% tres, y el 6,3% cuatro. En estas cinco primeras categorías están el 80,1% de los casos de la distribución, y el 20% restante se "reparte" entre el cinco y 15 libros que declara leer la persona que más lee (esto es, queda distribuida en un elevado número de categorías).

Realizar una categorización con *percentiles iguales* implicaría considerar los últimos valores de la distribución en una misma categoría, donde quedaría recogido el 20% de las respuestas (parte izquierda de la tabla 8.6, última categoría). Ahora bien, de haber creado intervalos de igual amplitud se obtendrían las categorías que se muestran en

86. En este ejemplo ambas categorizaciones coinciden en los dos primeros *cortes*, pero no así en el tercero.

((v05) Número de horas libres que dispone a la semana de ocio o diversión					
		Frecuencia	**Porcentaje**	**Porcentaje válido**	**Porcentaje acumulado**
Válidos	7	2	1,0	1,1	1,1
	8	2	1,0	1,1	2,2
	10	10	5,2	5,4	7,5
	12	2	1,0	1,1	8,6
	14	4	2,1	2,2	10,8
	15	3	1,6	1,6	12,4
	16	3	1,6	1,6	14,0
	18	2	1,0	1,1	15,1
	20	26	13,6	14,0	29,0
	24	4	2,1	2,2	31,2
	25	12	6,3	6,5	37,6
	30	10	5,2	5,4	43,0
	35	8	4,2	4,3	47,3
	36	2	1,0	1,1	48,4
	40	26	13,6	14,0	62,4
	43	2	1,0	1,1	63,4
	44	4	2,1	2,2	65,6
	45	4	2,1	2,2	67,7
	48	6	3,1	3,2	71,0
	50	12	6,3	6,5	77,4
	55	4	2,1	2,2	79,6
	58	2	1,0	1,1	80,6
	60	4	2,1	2,2	82,8
	61	2	1,0	1,1	83,9
	62	2	1,0	1,1	84,9
	63	2	1,0	1,1	86,0
	68	2	1,0	1,1	87,1
	70	6	3,1	3,2	90,3
	72	8	4,2	4,3	94,6
	78	4	2,1	2,2	96,8
	80	6	3,1	3,2	100,0
	Total	186	97,4	100,0	
Perdidos	No responde	5	2,6		
Total		191	100,0		

Tabla 8.4. Proceso de categorización.

(v06) Número de libros, relacionados con los estudios, leídos en el último año

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Ninguno	39	20,4	20,4	20,4
	1	48	25,1	25,1	45,5
	2	28	14,7	14,7	60,2
	3	26	13,6	13,6	73,8
	4	12	6,3	6,3	80,1
	5	8	4,2	4,2	84,3
	6	12	6,3	6,3	90,6
	7	4	2,1	2,1	92,7
	8	4	2,1	2,1	94,8
	10	6	3,1	3,1	97,9
	11	2	1,0	1,0	99,0
	15	2	1,0	1,0	100,0
	Total	191	100,0	100,0	

Tabla 8.5. Distribución de una variable muy asimétrica.

la parte derecha de la tabla 8.6; donde destaca una categoría (la primera) con un gran número de casos, y la última con muy pocos casos: cuatro entrevistados[87].

Aunque de las dos soluciones nos parece mejor la situada a la izquierda, es indudable que hubiera sido mejor *separar* los que no leen libros de aquellos que leen uno. Esto puede hacerse cambiando los valores del cuadro de diálogo *Crear puntos de corte*[88] (figura 8.6) o, más sencillo, empleando una recodificación manual; aspecto que veremos en los apartados 4 y 5.

Al igual que en el ejemplo anterior recomendamos situar la variable categorizada cerca de la variable original, con el fin de tener claro el origen de la *nueva* varia-

87. Este bajo número de casos presenta algunas complicaciones a la hora de relacionar variables en tablas de contingencia, como veremos en el apartado 2 del capítulo IX.

88. A continuación se presenta la distribución obtenida manteniendo el mismo número de puntos de corte (3), y aumentando la *anchura* hasta 4. En este caso la información proporcionada por el programa –al pulsar el tabulador– es 0 como posición del primer corte, y 8 como último punto de corte:

Menos de cero	20,4%
Entre 1 y 4	59,7%
Entre 5 y 8	14,7%
Entre 9 y 15	5,2%

Obsérvese el gran número de casos de la segunda categoría (59,7%, 74 entrevistados), y el escaso tamaño de la última (5,2%, 10 entrevistados).

(v06) Número de libros, relacionados con los estudios, leídos en el último año (Categorizada)			
Percentiles iguales		**Intervalos de igual amplitud[89]**	
Categoría	**%**	**Categoría**	**%**
Ninguno y uno	45,5%	Menos de uno	45,5%
Dos	14,7%	Entre 2 y 6	38,7%
Tres y cuatro	19,9%	Entre 7 y 10	13,6%
Entre cinco y quince	19,9%	Entre 11 y 15	2,1%
Total	191	Total	191

Tabla 8.6. Resultado de Categorizar Variables: comparación entre los dos ambos procedimientos en una variable asimétrica.

ble. Al final del apartado 8.2 señalamos una forma de *trasladar* la nueva variable del final del archivo de datos a un lugar cercano a la variable de procedencia. Existe otra forma de realizar esta operación que consiste en crear la nueva variable antes de realizar la operación correspondiente. Así, antes de proceder con la categorización es necesario *insertar* una nueva variable entre v05 y v06[90] y definir esta variable (vacía) con el nombre correspondiente (v05_cat, por ejemplo). Cuando se proceda con la categorización (o recodificación, cálculos, etc.) se indicará al programa este nombre como *variable destino*, y el programa preguntará si desea reemplazar la variable existente. Tras pulsar *Aceptar* aparecerán en esta nueva variable los valores correspondientes[91].

El proceso de categorización termina incluyendo la información de esta nueva variable en el libro de códigos y, por supuesto, guardando el archivo de datos para conservar así las nuevas variables categorizadas. Obsérvese la distribución de las variables v05_ca_1, v05_ca_2 y v06_cat en el libro de códigos final (*materiales complementarios*).

Buscando *fijar* los conocimientos aprendidos en esta sección, antes de considerar nuevos contenidos, proponemos dos ejercicios utilizando la investigación sobre *Vida Cotidiana*. Se trata de categorizar dos variables: Pensando en los amigos más íntimos, desde hace cuántos años conoce al amigo que sea más antiguo? (pregunta 34a, variable b66). La segunda recoge información sobre el número de amigos de verdad de los entrevistados (pregunta 34b, variable b68).

89. Las instrucciones introducidas en el cuadro de diálogo "crear puntos de corte" (figura 8.6) han sido: Número de puntos de corte 3, y 1 como posición del primer corte. Tras pulsar el tabulador el programa presenta la siguiente información: amplitud 4,67, y 10 como último punto de corte.
90. En el caso de la categorización de la variable v05. Recuérdese que en el penúltimo párrafo de la sección anterior se explicó como insertar variables en el archivo de datos.
91. Conviene señalar que esta forma de trasladar las nuevas variables puede utilizarse en todos los procedimientos presentados en este capítulo, excepto en la *recodificación automática*.

4. Recodificar en las mismas variables

Hasta ahora se han creado nuevas variables con la transformación de otras, dejando que sea el propio programa el que realice el proceso. Los procedimientos que se presentan a continuación tienen en común que es el propio investigador el que decide el criterio para realizar la transformación de las variables, y se diferencian en el tipo de transformación efectuado. Dedicaremos este apartado y el siguiente a la unión de categorías, en el 6 se explicará cómo realizar cálculos con las variables, y en el 7 se mostrará el proceso de creación de nuevas variables uniendo determinados valores de otras variables.

En numerosas ocasiones las variables utilizadas en una investigación precisan de transformaciones que consisten en agrupar varias categorías en una sola, eliminación de determinados valores, etc.; cambios que serán expuestos en esta sección y en la siguiente. Considerando las frecuencias presentadas en la tabla 7.1, por ejemplo, es necesario interpretar con sumo cuidado las opciones que han sido elegidas por pocos entrevistados, no pudiendo sacar conclusiones significativas en aquellas que han sido elegidas por menos del 5% de los entrevistados[92]. En la tabla 7.1 esto implica referirse con sumo cuidado a dos de las opciones presentadas, concretamente "aburrirme" y "pensar meditar".

Considerando este hecho, ¿qué interpretación puede realizarse de las frecuencias de la pregunta 1 (variable v01) propuesta en el primer ejercicio del capítulo 7 (ver *materiales complementarios*)? El gran número de opciones de respuesta de esta pregunta, unido a las escasas elecciones recibidas por algunas, recomienda agrupar las categorías que no alcanzan el 5% con otras de temática similar o, en su defecto, en una opción conjunta denominada "otras". Considerando ambos criterios se ha elaborado una categoría que agrupa las respuestas *ir de excursión* (4,2%) e *ir al monte* (2,1%); mientras que el resto de respuestas forman parte de la categoría denominada *otras*. Concretamente *ir al teatro* (1,0%), *ir a conciertos* (2,1%), *leer libros* (1,0%), *otras* (3,1%), *ninguna en particular* (1,0%), *quedar con amigos* (1,0%) y *quedar con el novio* (1,0%)[93]. Téngase en

92. De momento basta con que nos quedemos con este dato; que hay que tener sumo cuidado con aquellas categorías con valores porcentuales inferiores al 5%, es decir, que han sido elegidas por menos del 5% de los entrevistados. A los interesados en lograr una explicación de este hecho recomendamos la lectura de Rodríguez Osuna et al (1991: 58–60) y Díaz de Rada (1999: 96–99).

93. Obsérvese que las tres últimas respuestas de esta pregunta han sido *elaboradas* por el investigador al recoger las respuestas proporcionadas a la *categoría abierta* de la pregunta 1. En aquel caso se tomaron las respuestas literales del entrevistado, y ahora debemos agruparla con otras categorías. Si la pregunta utilizada está bien diseñada se espera que la opción "otras" proporcione muy pocas *nuevas respuestas*, por lo que la mayor parte de las veces tan sólo se anota la opción *otra*"; esto es, sin detallar a que otras situaciones se refiere (siempre que –como señala el estándar de

cuenta que estos porcentajes tan bajos están indicando que estas opciones han sido elegidas por 2 o tres personas, no pudiendo sacar resultados representativos de un número de entrevistados tan bajo. Para este propósito utilizaremos la recodificación.

El objetivo es reducir las categorías de una variable buscando *eliminar* las respuestas con escasas elecciones pero conservando –a la vez– el máximo número de categorías. Se busca una opción intermedia entre ambas situaciones.

Para recodificar hay que considerar el número de casos y la similitud temática entre categorías; tratando de unir categorías similares con pocos casos, y mantener las categorías con un número suficiente de respuestas. Para explicar la recodificación, que no es otra cosa que la unión de categorías, nos serviremos –como en otras ocasiones– de un ejemplo, concretamente la pregunta 16 (v43) sobre la situación de residencia de los estudiantes de sociología y sus amigos. El objetivo de esta pregunta era conocer el porcentaje de estudiantes que vivían con su familia de origen (padres, padre o madre, un progenitor y un abuelo), en piso con compañeros, o con su propia familia. Con este fin se utilizó una pregunta con 3 categorías de respuesta[94], a la que se añadió una cuarta para que el entrevistado apuntara su situación cuando no estuviera dentro de las anteriores.

Las respuestas se muestran en la tabla 8.7, que desvela algunas incoherencias en determinadas respuestas:

- La opción 4 (vivir con padre o madre), ¿no es vivir con la familia de origen?
- La opción 7 (vivir con madre y hermanos), ¿no es vivir con la familia de origen?
- La opción 9 (vivir con madre y abuela), ¿no es vivir con la familia de origen?

---

calidad de ANEIMO [2000: 251]– esta categoría no supere el 10% de las respuestas obtenidas). No obstante, y por motivos didácticos, en el tercer capítulo codificamos todas las respuestas proporcionadas por la pregunta.

De forma diferente procedimos en la pregunta 21, al codificar como *otros* a todos los entrevistados que no habían estudiado en la Universidad Pública de Navarra, en la Universidad de Navarra, y en la Universidad Nacional de Educación a Distancia. El número de universidades presenta una gran variación, pero el hecho que los cuestionarios hayan sido respondidos por amigos de estudiantes de Sociología limita notablemente la posibilidad de conseguir un gran número de personas que estudien en universidades diferentes a las apuntadas. Por este motivo no se recoge el nombre de la universidad, ya que al ser muy pocos los entrevistados de otros centros (concretamente 10 personas) no es posible analizar separadamente los estudiantes de cada centro (el escaso número de personas impide conseguir resultados representativos).

Utilizaremos una cita literal de J.I. Wert para indicar la forma más adecuada de proceder para el *cierre* este tipo de preguntas: "...reducir a unas categorías sintéticas las posibilidades de respuesta en función de una 'muestra de la muestra' (normalmente entre un 20% y un 30% del total suele ser suficiente) de cuyas respuestas literales se puede extraer un número limitado de categorías de respuesta, digamos diez o doce... Este proceso de agrupamiento y síntesis es la confección del llamado *libro de códigos*..." (Wert, 1996: 58).

94. Pregunta 16: ¿Con quién vives?
– CON MIS PADRES 1
– CON AMIGOS EN UN PISO COMPARTIDO 2
– CON MI PAREJA 3
– OTRAS SITUACIONES (apuntar) _____

Dicho de otro modo, será necesario agrupar los valores 4, 7 y 9 con el 1 (con padres, familia de origen), y el 5 con el 3 (familia propia). El resto de valores no citados, el 6 (vivir con hermanos), el 8 (vivir con tíos y primos) y el 10 (en residencia y colegio mayor) podrán ser agrupados en una nueva categoría definida como *otras situaciones*.

(v43) Vive con

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Con padres (1)	128	67,0	67,0	67,0
	Con amigos en piso compartido (2)	28	14,7	14,7	81,7
	Con pareja (3)	10	5,2	5,2	86,9
	Otras situaciones: con padre ó madre (4)	4	2,1	2,1	89,0
	Otras situaciones: en familia propia(5)	6	3,1	3,1	92,1
	Otras situaciones: con hermanos (6)	2	1,0	1,0	93,2
	Otras situaciones: con madre y hermanos (7)	2	1,0	1,0	94,2
	Otras situaciones: con tíos y primos (8)	1	,5	,5	94,8
	Otras situaciones: madre y abuela (9)	4	2,1	2,1	96,9
	Otras situaciones: en residencia (10)	6	3,1	3,1	100,0
	Total	191	100,0	100,0	

Tabla 8.7. Situación de residencia. Vive con:

El SPSS dispone de dos tipos de recodificaciones, recodificar en las mismas o en distintas variables. Comenzaremos por el primero, puesto que es el más sencillo, para ir posteriormente complicando la exposición. Marcando *Transformar⇒Recodificar en las mismas variables* aparece el cuadro de diálogo de la figura 8.9. Pulsando el botón *Valores antiguos y nuevos* se presenta el cuadro de diálogo 8.10 donde se produce la reconversión de unos valores en otros. La parte izquierda está dedicada a los valores antiguos, y la derecha a los valores nuevos.

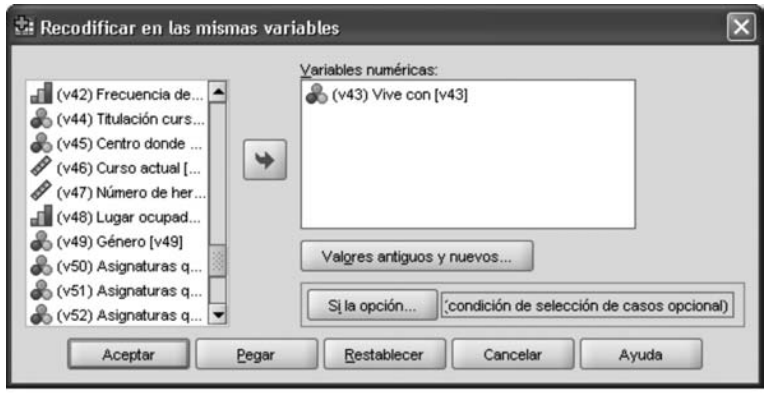


Figura 8.9. Cuadro de diálogo Recodificar en las mismas variables.

El proceso seguido ha sido el siguiente: se introduce el 4 en el recuadro superior izquierdo (debajo de valor antiguo), el 1 en el recuadro superior derecho (debajo de valor nuevo), y se pulsa el botón *Añadir* situado en el centro. En esta ventana parece escrito *4→1*, que indica que el valor 4 se convierte en 1. Así se ha procedido hasta colocar todos los valores, tal y como se muestra en la figura 8.10. De hecho, en esta figura únicamente falta pulsar *Añadir* para que el 10 se convierta en 4.

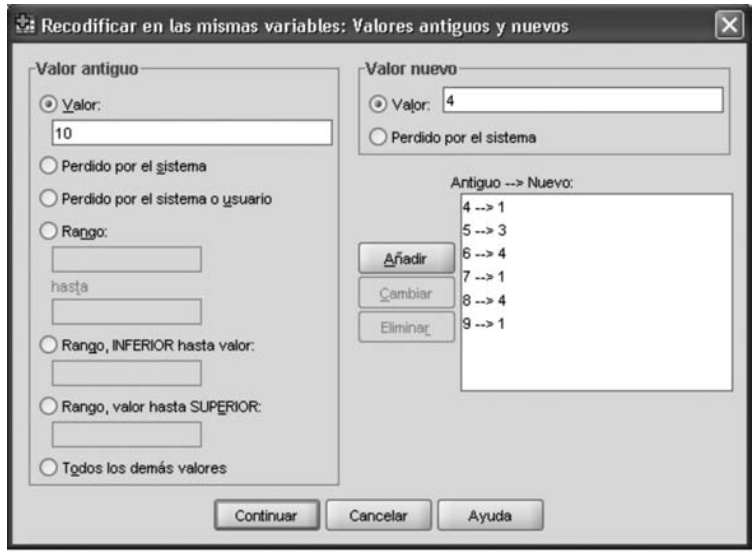


Figura 8.10. Cuadro de diálogo Recodificar en las mismas variables: valores antiguos y nuevos.

Pulsando *Continuar*[95] (en figura 8.10) y *Aceptar* (en figura 8.9) se lleva a cabo la sustitución de unos valores por otros en la ventana de datos. Tras cambiar la etiqueta del valor 4[96] será necesario solicitar de nuevo las frecuencias para valorar el resultado de esta acción; que se muestra en la tabla 8.8. No debe olvidarse introducir esta modificación en el libro de códigos.

(v43) Vive con					
		Frecuencia	**Porcentaje**	**Porcentaje válido**	**Porcentaje acumulado**
Válidos	Con padres	138	72,3	72,3	72,3
	Con amigos en piso compartido	28	14,7	14,7	86,9
	Con pareja	16	8,4	8,4	95,3
	Otras situaciones	9	4,7	4,7	100,0
	Total	191	100,0	100,0	

Tabla 8.8. Situación de residencia recodificada.

Otra opción más compleja que posibilita este paquete estadístico es realizar la recodificación de la variable si se cumplen determinadas condiciones. Por ejemplo podemos indicar que únicamente realice esa categorización si la persona entrevistada es una mujer, un estudiante de primero, etc. Para ello será preciso marcar, dentro del cuadro de diálogo *Recodificar en las mismas variables* (figura 8.9) el botón *Si la opción...* y establecer la condición que deseamos, tal y como se muestra en la figura 8.11. En este ejemplo se ha seleccionado *Incluir si el caso satisface la condición...* para trasladar seguidamente la V49 a la ventana superior, y añadir los símbolos = y 2. En este caso la recategorización efectuada únicamente afectara a las respuestas dadas por las mujeres (valor dos en la variable v49, sexo), y la variable aparecerá sin recodificar cuando se analicen las opiniones de los hombres. Como ahora no interesa este procedimiento, se procederá a su desactivación pulsando la opción *Incluir todos los casos* en el cuadro de diálogo de la figura 8.11.

95. Si se pulsa *Continuar* antes de *Añadir* el programa advierte que no se ha terminado con la recodificación, indicando que "Se perderán todas las operaciones pendientes de Añadir o Cambiar". Será preciso seleccionar *Cancelar* para que desaparezca este cuadro de diálogo y así pulsar, a continuación, el botón *Añadir*.
96. Recordar sección 3.3, concretamente la parte dedicada a las *etiquetas de valor* (*valores*).

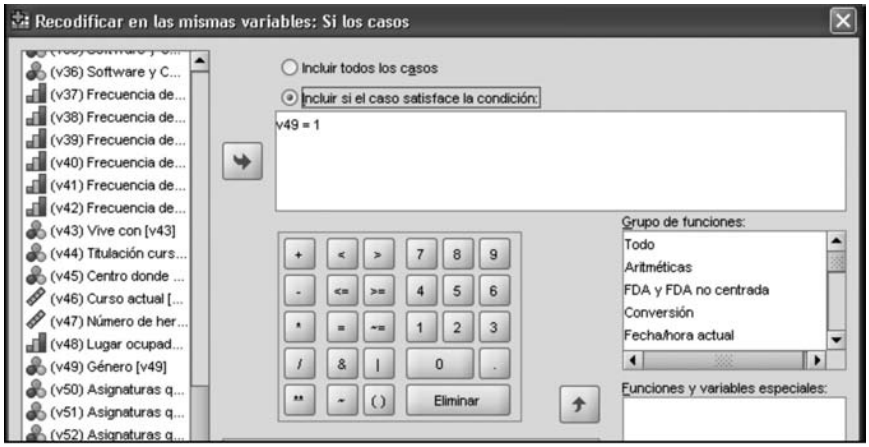


Figura 8.11. Cuadro de diálogo Recodificar en las mismas variables: Si la opción...

La siguiente vez que se utilice el procedimiento *Recodificar* (en las mismas o distintas variables) aparecerá la última recodificación efectuada en la actual sesión de trabajo. Será necesario pulsar el botón *Reestablecer* para proceder con la nueva recodificación.

Deseamos insistir en la necesidad de realizar estas modificaciones con el esmero y la preocupación constante de no alterar la información proporcionada por los datos, cuidando al máximo la objetividad del proceso y considerando en todo momento la afinidad teórica entre categorías. El reagrupar categorías implica una pérdida de información de modo que SIEMPRE deberemos hacer este proceso con sumo cuidado. Antes de cualquier reagrupación debe valorarse hasta que punto podemos permitir perder determinada información y qué ventajas se consiguen con ello. Además, debe tenerse en cuenta que una vez agrupadas las categorías no es posible volver atrás. O dicho de otro modo, tras la recodificación la variable v43 tiene tan sólo cuatro categorías, y no es posible volver a obtener con ella la tabla 8.7.

Recordemos que la agrupación de variables nominales se realiza –como hemos visto– agrupando las categorías con pocas elecciones formando una categoría común denominada *otras*, o bien agrupando categorías *semejantes*. En variables ordinales suelen agruparse las categorías contiguas, aquellas que están próximas dentro de la "intensidad gradual" de la variable (Ruiz Maya et al, 1990: 183).

Fijar los conocimientos aprendidos es el objetivo de toda actividad docente. Con el fin de consolidar lo aprendido recomendamos –utilizando la investigación sobre *Vida Cotidiana*– solicitar las frecuencias de la variable estado civil (e12) y realizar las modificaciones pertinentes con el fin solventar los problemas señalados en el segundo y tercer párrafo de esta sección. Realizar una recodificación de la variable edad (e10) en cuatro grupos: de 18 a 29 años, de 30 a 44, de 45 a 64, y 65 y más años.

5. Recodificar en distintas variables

Existe la posibilidad de volver a obtener la distribución vista en la tabla 8.7 realizando la recodificación en distintas variables; esto es, dejar la variable original y llevar a cabo la recodificación en una copia de ésta. Utilizando la secuencia *Transformar⇒Recodificar en distintas variables* aparece el cuadro de diálogo de la figura 8.12 que permite *duplicar* una variable para realizar sobre ésta las transformaciones pertinentes. De este modo la variable original siempre tiene las respuestas dadas por los entrevistados.

Tras seleccionar la v43, por seguir con el mismo ejemplo del apartado anterior, será necesario elegir el nombre de la nueva variable (en este caso v43recod) y su etiqueta, para seguidamente pulsar el botón *Cambiar*. En este momento la nueva variable se traslada a la ventana del centro (figura 8.13), con un símbolo que indica que los valores de v43 pasarán a v43recod (v43→v43recod).

A continuación se pulsa el botón *Valores antiguos y nuevos*, y se procede exactamente igual que en la sección anterior, si bien en esta recodificación es necesario incluir todos los valores de las variables, incluso los que no sufren ninguna transformación. Esto explica, en la ventana derecha de la figura 8.14, los símbolos 1→1, 2→2, 3→3.

Figura 8.12. Cuadro de diálogo Recodificar en distintas variables.

Figura 8.13. Cuadro de diálogo Recodificar en distintas variables: v40recod.

Pulsando *Continuar* en la figura 8.14 y Aceptar (en la 8.13) se lleva a cabo la creación de la variable v43recod[97]. Será necesario solicitar las frecuencias de esta variable para valorar el resultado del proceso efectuado, y posteriormente introducir su información en el libro de códigos. Recomendamos, como en anteriores ocasiones, colocar esta variable cercana a la v43.

En este apartado y en el anterior (8.4) se han expuesto diversos procesos de recodificación para que cada investigador seleccione el que más le interese. Antes de terminar con este aspecto algunos consejos sobre las ventajas e inconvenientes de cada uno: la recodificación en distintas variables tiene a su favor que siempre conserva los valores originales, y quizás en determinados momentos nos interese diferenciar los que *viven con madre y abuela* de aquellos que *viven con hermanos* (por seguir con el ejemplo utilizado). Este proceso tiene el inconveniente que al duplicar variables aumenta el tamaño del archivo de datos, perdiendo velocidad de procesamiento y, lo que es más importante, generando que en determinados momentos no se tenga claro la procedencia y la diferencia entre determinadas variables. Para evitar este problema recomendamos nombrar a las nuevas variables con el nombre de la variable original, añadiendo la terminación "-bis", "-dos", etc.

97. Este procedimiento se utilizó en el archivo "Consumo Navarra" para crear la variable "v0036_re"; que no es otra cosa que la edad (v0036) recodificada en 5 grupos: 16-25 años, 26-35 años, 36-45 años, 46-55 años, 56-65 años.

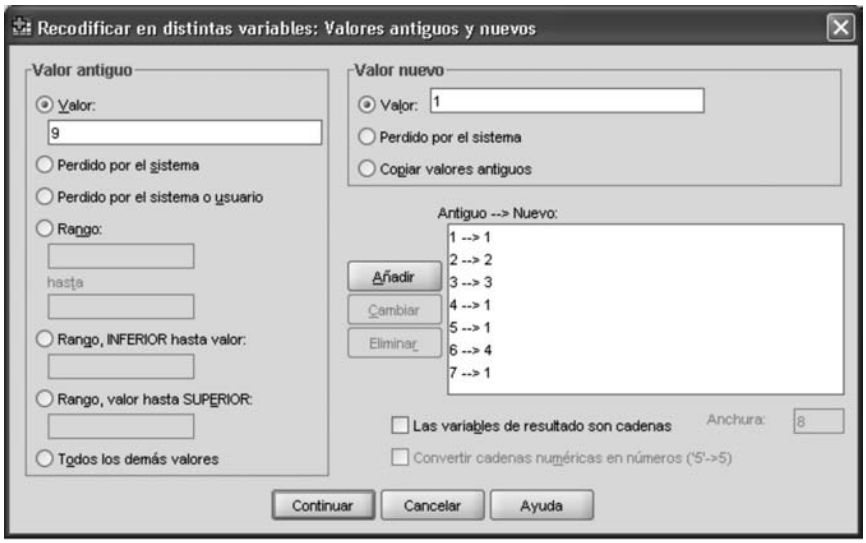


Figura 8.14. Cuadro de diálogo Recodificar en distintas variables: valores antiguos y nuevos.

Por otro lado, recodificar en las mismas variables permite trabajar con los nombres de variables que el investigador está familiarizado, evitando duplicidades y grandes tamaños del fichero de datos. Tiene el enorme inconveniente que no es posible volver a la versión original de los datos recogidos puesto que las versiones Windows del SPSS modifican el archivo de datos al realizar las recodificaciones en la propia matriz de datos, a diferencia de las versiones anteriores que no producían ningún cambio en el fichero de datos. Es por ello por lo que aconsejamos guardar SIEMPRE una copia de los datos originales fuera del ordenador.

Con el fin de fijar los conocimientos aprendidos en esta sección, antes de considerar nuevos contenidos, realizar –en la investigación sobre *Vida Cotidiana*– una recodificación de la variable edad (e10) en una nueva variable que se llame *edad*, considerando los cuatro grupos definidos en la sección anterior. Un vistazo al nivel de estudios del entrevistado (e21) desvela una excesiva dispersión de la información recogida, por lo que recomendamos reagrupar algunas categorías con el fin de establecer cuatro divisiones: sin estudios (donde se agrupan los que presentan menos de estudios primarios, sepan o no leer), estudios primarios, estudios segundarios (FP1, PF2, Bachiller Elemental y Bachiller Superior) y estudios superiores (estudios de escuela universitaria y de universidad).

6. Cálculos y operaciones: procedimiento calcular variable

En los ejercicios del capítulo siete (ver *materiales complementarios*) se preguntó por el número medio de libros no relacionados con sus estudios que los entrevistados habían leído en el último año. Ahora nos gustaría conocer el número total de libros leídos el pasado año; pero tras analizar el cuestionario (apartado 2.6) vemos que no hay ninguna variable donde obtener esta información. Sin embargo, una lectura más exhaustiva descubre que puede obtenerse agregando, al número de libros leídos no relacionados con sus estudios (v08), los libros leídos relacionados con sus estudios (v06). Es decir, deberíamos sumar las variables v08 y v06.

Esta situación constituye un excelente ejemplo para utilizar el procedimiento calcular. Seleccionado *Transformar⇒Calcular Variable* aparece el cuadro de diálogo mostrado en la figura 8.15. Tras colocar el nombre de la variable destino en la ventana superior izquierda, se construye la expresión numérica v06 + v08 en la ventana de la derecha. Posteriormente se indicará el tipo de variable y la etiqueta pulsando en el botón correspondiente (Figura 8.15).

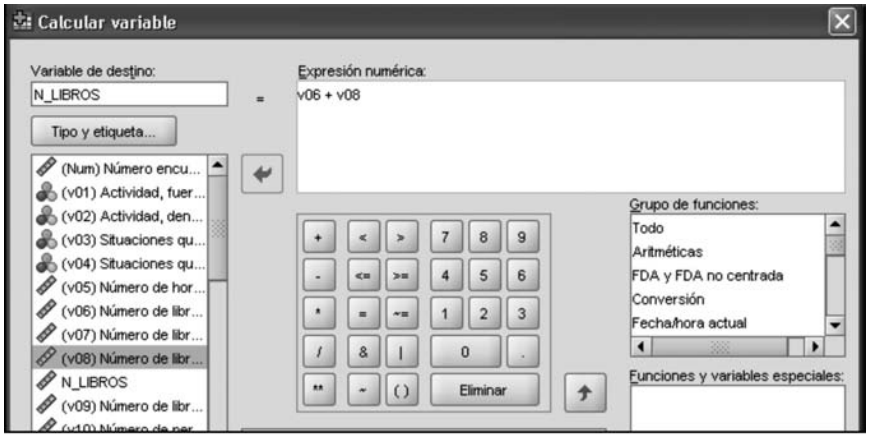


Figura 8.15. Cuadro de diálogo Calcular: suma de variables v06 + v08.

Pulsando Aceptar (en figura 8.15) se añade la nueva variable a la ventana de datos. El investigador deberá, en este momento, comprobar que la nueva variable es correcta. Para ello basta con realizar manualmente la suma de v06 y v08 en una muestra de entrevistados del archivo de datos con el fin de comprobar que este resultado (manual) coincide con la nueva variable. SIEMPRE debe hacerse esta comprobación.

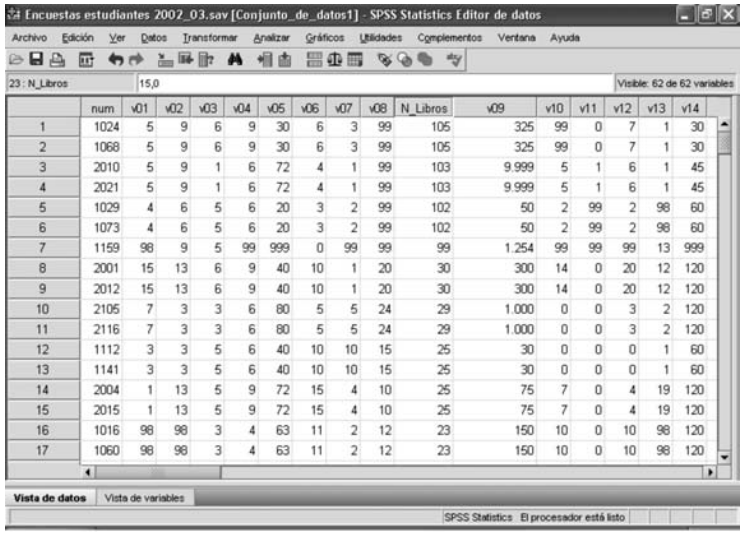
Tras la verificación, la información de la nueva variable deberá ser incluida en el libro de códigos final. Las frecuencias de la variable creada en la figura 8.15 se muestran en la tabla 8.9.

N_LIBROS. Número total de libros leídos					
		Frecuencia	**Porcentaje**	**Porcentaje válido**	**Porcentaje acumulado**
Válidos	,00	22	11,5	11,5	11,5
	1,00	12	6,3	6,3	17,8
	2,00	28	14,7	14,7	32,5
	3,00	16	8,4	8,4	40,8
	4,00	26	13,6	13,6	54,5
	5,00	4	2,1	2,1	56,5
	6,00	14	7,3	7,3	63,9
	7,00	12	6,3	6,3	70,2
	8,00	6	3,1	3,1	73,3
	9,00	6	3,1	3,1	76,4
	10,00	8	4,2	4,2	80,6
	11,00	2	1,0	1,0	81,7
	12,00	6	3,1	3,1	84,8
	13,00	4	2,1	2,1	86,9
	14,00	2	1,0	1,0	88,0
	15,00	2	1,0	1,0	89,0
	17,00	2	1,0	1,0	90,1
	21,00	2	1,0	1,0	91,1
	23,00	2	1,0	1,0	92,1
	25,00	4	2,1	2,1	94,2
	29,00	2	1,0	1,0	95,3
	30,00	2	1,0	1,0	96,3
	99,00	1	,5	,5	96,9
	102,00	2	1,0	1,0	97,9
	103,00	2	1,0	1,0	99,0
	105,00	2	1,0	1,0	100,0
	Total	191	100,0	100,0	

Tabla 8.9. Número total de libros leídos en el último año: libros relacionados con sus estudios (v06) + libros no relacionados con sus estudios (v08).

Como puede comprobarse siete entrevistados han leído más de 90 libros en los últimos meses: una persona ha leído 99, dos han leído 102 libros, otras dos 103, y otras tantas 105. ¿Es posible? Será interesante analizar estos seis casos, ver quiénes son, y para ello utilizaremos algunos procedimientos vistos en el sexto capítulo (lo que nos servirá de repaso). La forma más efectiva para ver quiénes son estos entrevistados es realizar una ordenación de archivo de datos en función de los valores de la nueva variable creada. Con este fin utilizaremos el menú *Datos⇒Ordenar casos*, colocando N_Libros en el espacio correspondiente, y solicitando un *orden de clasificación descendente* (ver figura 6.6). Otra posibilidad es –seleccionada la variable N_Libros– buscar los valores 102, 103 y 105 con el procedimiento *Edición⇒Buscar*.

Lógicamente, al contar con un archivo de datos relativamente pequeño es posible *moverse* por la pantalla de datos hasta localizar lo que estamos buscando; pero ¿qué ocurriera si tuviéramos –por ejemplo– 5.000 entrevistados?, que es la situación más habitual. En este caso la localización visionando el archivo de datos resulta más complicada. Para afrontar esta situación recomendamos utilizar los procedimientos *Buscar* y *Ordenar Casos*, si bien consideramos que el último cumple mejor la misión encomendada en esta ocasión. En la figura 8.16 se muestra el resultado de la ordenación descendente considerando los valores de la variable N_libros[98].



	num	v01	v02	v03	v04	v05	v06	v07	v08	N_Libros	v09	v10	v11	v12	v13	v14
1	1024	5	9	6	9	30	6	3	99	105	325	99	0	7	1	30
2	1068	5	9	6	9	30	6	3	99	105	325	99	0	7	1	30
3	2010	5	9	1	6	72	4	1	99	103	9.999	5	1	6	1	45
4	2021	5	9	1	6	72	4	1	99	103	9.999	5	1	6	1	45
5	1029	4	6	5	6	20	3	2	99	102	50	2	99	2	98	60
6	1073	4	6	5	6	20	3	2	99	102	50	2	99	2	98	60
7	1159	98	9	5	99	999	0	99	99	99	1.254	99	99	99	13	999
8	2001	15	13	6	9	40	10	1	20	30	300	14	0	20	12	120
9	2012	15	13	6	9	40	10	1	20	30	300	14	0	20	12	120
10	2105	7	3	3	6	80	5	5	24	29	1.000	0	0	3	2	120
11	2116	7	3	3	6	80	5	5	24	29	1.000	0	0	3	2	120
12	1112	3	3	5	6	40	10	10	15	25	30	0	0	0	1	60
13	1141	3	3	5	6	40	10	10	15	25	30	0	0	0	1	60
14	2004	1	13	5	9	72	15	4	10	25	75	7	0	4	19	120
15	2015	1	13	5	9	72	15	4	10	25	75	7	0	4	19	120
16	1016	98	98	3	4	63	11	2	12	23	150	10	0	10	98	120
17	1060	98	98	3	4	63	11	2	12	23	150	10	0	10	98	120

Figura 8.16. Ventana de datos para observar la ordenación *descendente* del archivo de datos considerando la variable N_libros.

98. Esta vista del archivo puede aprovecharse también para constatar que la nueva variable se ha construido correctamente, que la suma *manual* de v06 y v08 coincide con el valor de cada entrevistado en la variable N_Libros.

Un análisis exhaustivo de la figura 8.16 permite identificar claramente los motivos por los que seis entrevistados ha leído más de 100 libros en los últimos meses: este resultado se ha producido por una incorrecta definición del código *no responde* en la variable v08 (puesto que todo el mundo ha respondido v06). Existen dos formas de corregir este error:

- La primera se fundamenta en la asignación del valor 99 como valor perdido en la variable v08, proceso que debe hacerse antes de realizar el cálculo, antes de crear la variable N_libros.

N_LIBROS. Número total de libros leídos

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	,00	22	11,5	12,0	12,0
	1,00	12	6,3	6,5	18,5
	2,00	28	14,7	15,2	33,7
	3,00	16	8,4	8,7	42,4
	4,00	26	13,6	14,1	56,5
	5,00	4	2,1	2,2	58,7
	6,00	14	7,3	7,6	66,3
	7,00	12	6,3	6,5	72,8
	8,00	6	3,1	3,3	76,1
	9,00	6	3,1	3,3	79,3
	10,00	8	4,2	4,3	83,7
	11,00	2	1,0	1,1	84,8
	12,00	6	3,1	3,3	88,0
	13,00	4	2,1	2,2	90,2
	14,00	2	1,0	1,1	91,3
	15,00	2	1,0	1,1	92,4
	17,00	2	1,0	1,1	93,5
	21,00	2	1,0	1,1	94,6
	23,00	2	1,0	1,1	95,7
	25,00	4	2,1	2,2	97,8
	29,00	2	1,0	1,1	98,9
	30,00	2	1,0	1,1	100,0
	Total	184	96,3	100,0	
Perdidos	99,00	1	,5		
	102,00	2	1,0		
	103,00	2	1,0		
	105,00	2	1,0		
	Total	7	3,7		
Total		191	100,0		

Tabla 8.10. Número total de libros leídos en el último año, eliminados los valores "no posibles".

- La segunda considera la situación en la que nos encontramos, es decir, una vez creada la variable N_libros y detectardo el error. En este caso se definirían los valores perdidos en la nueva variable, y serían todos aquellos que superan el valor 99. Para conseguir tal propósito se abre el cuadro de diálogo *Valores perdidos* y, tras seleccionar la opción *Rango más un valor perdido discreto opcional* se escribe el rango de los valores *incoherentes* observados en la tabla 8.9; concretamente del 99 al 105. En la tabla 8.10 se muestra la tabla de frecuencias obtenida.
  Obsérvese que se pierden 7 casos, siete entrevistados que no han respondido una de las preguntas y que por lo tanto no pueden ser considerados en los análisis. No es que estas personas no lean libros no relacionados con sus estudios, es que no han respondido esta pregunta, hecho que plantea dificultades a la hora de considerar el número total de libros leídos. Por la falta de información proporcionada consideramos que es más adecuado eliminarlos de nuestros análisis.

Para terminar, será conveniente reducir el número de categorías de esta variable, utilizando la recodificación o cualquiera de los procedimientos empleados anteriormente. En la tabla 8.11 se presenta esta variable en seis categorías, lo que facilita notablemente su interpretación. Efectuada esta transformación, es el momento de considerar quiénes son los colectivos que más y menos leen: ¿existe diferencia considerando el sexo, curso, titulación, etc.?

12,0%	de los entrevistados	= 0 libros
21,7%	" " "	= 1-2 libros
22,8%	" " "	= 3-4 libros
16,3%	" " "	= 5, 6 y 7 libros
10,9%	" " "	= 8, 9 y 10 libros
16,3%	" " "	= más de 10.

Tabla 8.11. Número de libros leídos en el último año (variable N_libros reducida a seis categorías).

Con el fin de fijar los conocimientos aprendidos en esta sección recomendamos –utilizando el archivo *Vida Cotidiana*– elaborar una variable que muestre la diferencia entre el número de horas trabajadas y el número de horas que le gustaría trabajar, aspectos que se recogen en la pregunta 39 del cuestionario (variables b74 y b76). En la pregunta 37 se solicita del entrevistado que señale sus gastos en un conjunto

de cinco bienes y servicios (c65, c68, c71, c74 y c77). Posteriormente, en la variable d1, el entrevistador debe colocar la suma de todos esos valores. ¿Se han cometido errores en la respuesta de esta última?

7. Creación de nuevas variables uniendo valores en las variables de origen (contar valores dentro de los casos)

Otro de los procedimiento de transformación consiste en la creación de una variable seleccionando determinados valores de un conjunto de variables. Se verá mejor con un ejemplo que busca conocer el *número de asignaturas* que proponen libros de "lectura obligatoria". En la pregunta 7 (variables v50-v58) los entrevistados han señalado las asignaturas que proponen libros de lectura obligatoria. El análisis de tablas de frecuencias, o mejor aún la elaboración de una tabla multirespuesta categórica, puede proporcionar información sobre este tema; concretamente las asignaturas más señaladas por los estudiantes, aquellas sobre las que existe más acuerdo con que proponen libros de lectura obligatoria (tabla 8.12).

El problema es que en este momento no interesa tanto el nombre de las asignaturas, sino conocer el *número* de asignaturas que obligan a leer. Observando el libro de códigos (apartado 3.9) vemos que éstas están codificadas desde el 1 al 51, apareciendo también un valor 96 para los que no recuerdan, y el 97 por si entrevistado no reproduce bien el nombre de la asignatura.

El procedimiento *contar valores* creará una nueva variable con el número de asignaturas que, a juicio de cada entrevistado, proponen libros de lectura obligatoria. Para ello seleccionamos el menú *Transformar⇒Contar valores dentro de los casos*, obteniendo el cuadro de diálogo de la figura 8.17. Tras colocar el nombre y la etiqueta de la variable destino se eligen las variables de las que se obtendrá la información, en este caso de la v50 a la v58.

Pulsando el botón *Definir valores* se accede al cuadro de diálogo de la figura 8.18, donde se definen los valores a contar. Como hemos señalado que estas variables están codificadas del 1 al 97, pediremos al programa que considere (cuente) todos estos valores utilizando la opción *Rango* (parte izquierda de la figura 8.18). Al igual que en los procedimientos expuestos a lo largo de este capítulo, el botón *Añadir* colocará la expresión en la ventana de la derecha y, tras pulsar *Continuar* (figura 8.18) y *Aceptar* (figura 8.17) se añade la nueva variable (N_ asigna) a la ventana de datos. Como en anteriores ocasiones, hay que añadir esta variable, especificando la información que contiene, al libro de códigos final.

Group $Preg_7 Asignaturas que proponen libros de lectura obligatoria

Category label	Code	Count	Pct of Responses	Pct of Cases
Sociología General	1	72	19,7	55,4
Ciencia Política	2	18	4,9	13,8
Economía política	3	12	3,3	9,2
Historia Política y Social contemporánea	4	22	6,0	16,9
Estructura Social y Estructura Social de	6	18	4,9	13,8
Teoría Sociológica Clásica I	8	8	2,2	6,2
Teoría Sociológica I	9	24	6,6	18,5
Sistema Político Español	10	14	3,8	10,8
Métodos y Técnicas de investigación Soci	12	4	1,1	3,1
Historia de las Ideas Políticas	13	20	5,5	15,4
Sociología de la Comunicación y Opinión	16	24	6,6	18,5
Teoría Sociológica Clásica II	22	4	1,1	3,1
Antropología Social	24	8	2,2	6,2
Teoría de la Población	26	2,	5	1,5
Teoría Sociológica II	27	2,	5	1,5
Filosofía y Metodología de las Ciencias	28	2,	5	1,5
Política Social	31	2,	5	1,5
Ideologías Políticas Contemporáneas	35	2,	5	1,5
Procesos de Cambio Político en la Socied	36	2,	5	1,5
Novela postguerra	42	2,	5	1,5
Sociedad industrial	43	4	1,1	3,1
Otras 1 (No pertenecen al Plan de Estudi	50	42	11,5	32,3
Otras 2 (No pertenecen al Plan de Estudi	51	16	4,4	12,3
Otras 3 (No pertenecen al Plan de Estudi	52	14	3,8	10,8
Otras 3 (No pertenecen al Plan de Estudi	53	6	1,6	4,6
No queda claro. No reproduce bien el nom	97	22	6,0	16,9
Total responses		366	100,0	281,5

61 missing cases; 130 valid cases

Tabla 8.12. Frecuencias de las variables v50-v58 (Pregunta 7).

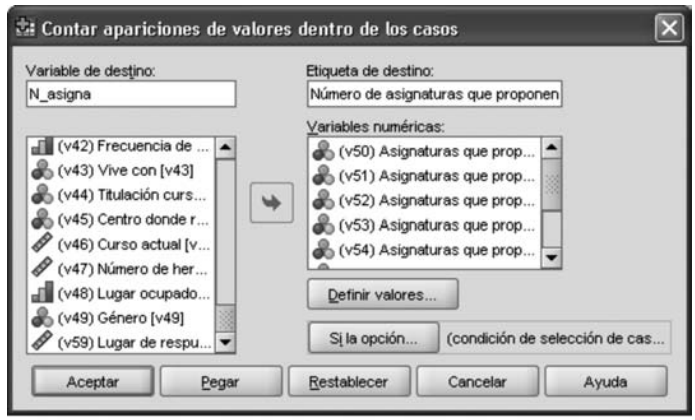


Figura 8.17. Cuadro de diálogo contar valores dentro de los casos.

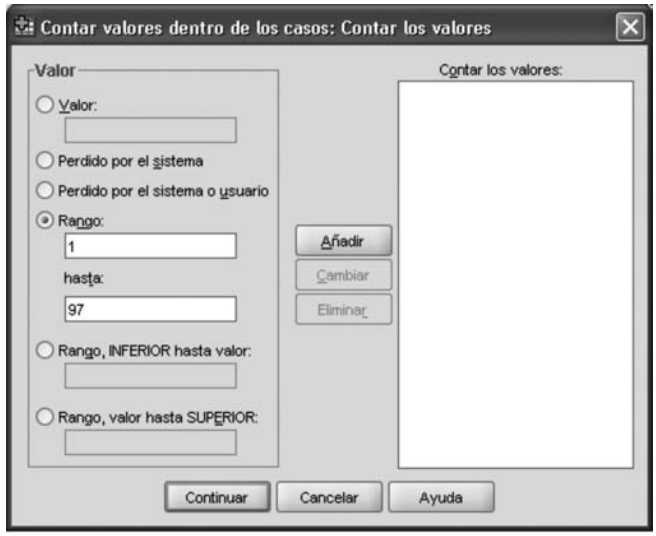
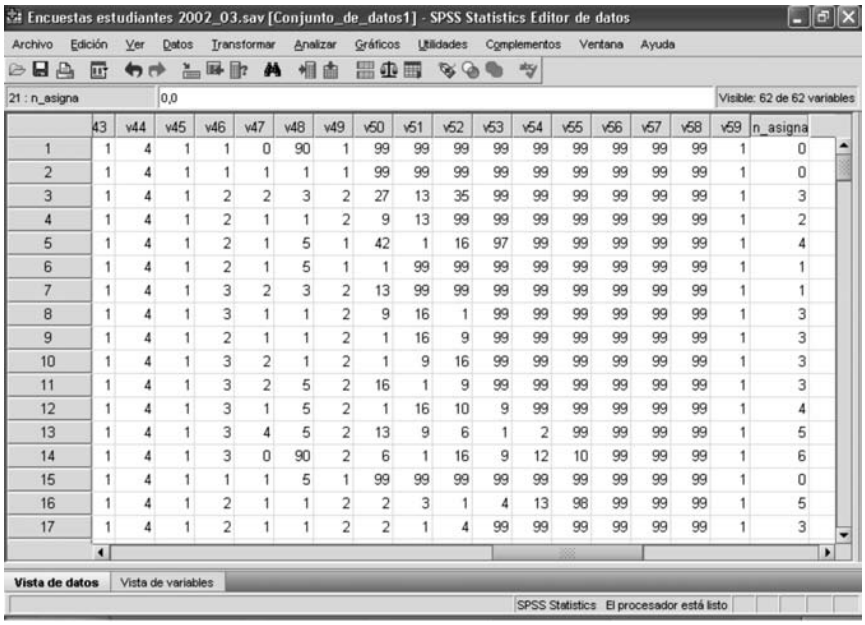


Figura 8.18. Cuadro de diálogo contar valores dentro...: Contar los valores.

En la figura 8.19 puede apreciarse como funciona el procedimiento contar valores dentro de los casos. El entrevistado número 3 (tercera línea) señaló que las asignaturas codificadas con los números 27, 13 y 35[99] proponen libros de lectura obliga-

99. El valor de cada una se muestra en el apartado 3.9. El número 27 se refiere a "Teoría Sociológica II", el 13 a "Historia de las Ideas Políticas", y el 35 a "Ideologías Políticas Contemporáneas".

toria (ver variables de la v50 a v58). Se trata de tres asignaturas, y por eso presenta el valor 3 en la variable "N_asigna". El entrevistado número cuatro eligió las asignaturas 9 y 13, y esto explica el valor 2 en la nueva variable. Por su parte el quinto encuestado (nº 5) considera que cuatro asignaturas proponen libros de lectura obligatoria, concretamente la 42, 1, 16 y 97, y por ese motivo tiene un 4 en la variable creada con el procedimiento contar valores.



	v43	v44	v45	v46	v47	v48	v49	v50	v51	v52	v53	v54	v55	v56	v57	v58	v59	n_asigna
1	1	4	1	1	0	90	1	99	99	99	99	99	99	99	99	99	1	0
2	1	4	1	1	1	1	1	99	99	99	99	99	99	99	99	99	1	0
3	1	4	1	2	2	3	2	27	13	35	99	99	99	99	99	99	1	3
4	1	4	1	2	1	1	2	9	13	99	99	99	99	99	99	99	1	2
5	1	4	1	2	1	5	1	42	1	16	97	99	99	99	99	99	1	4
6	1	4	1	2	1	5	1	1	99	99	99	99	99	99	99	99	1	1
7	1	4	1	3	2	3	2	13	99	99	99	99	99	99	99	99	1	1
8	1	4	1	3	1	1	2	9	16	1	99	99	99	99	99	99	1	3
9	1	4	1	2	1	1	2	1	16	9	99	99	99	99	99	99	1	3
10	1	4	1	3	2	1	2	1	9	16	99	99	99	99	99	99	1	3
11	1	4	1	3	2	5	2	16	1	9	99	99	99	99	99	99	1	3
12	1	4	1	3	1	5	2	1	16	10	9	99	99	99	99	99	1	4
13	1	4	1	3	4	5	2	13	9	6	1	2	99	99	99	99	1	5
14	1	4	1	3	0	90	2	6	1	16	9	12	10	99	99	99	1	6
15	1	4	1	1	1	5	1	99	99	99	99	99	99	99	99	99	1	0
16	1	4	1	2	1	1	2	2	3	1	4	13	98	99	99	99	1	5
17	1	4	1	2	1	1	2	2	1	4	99	99	99	99	99	99	1	3

Figura 8.19. Ventana de datos para observar el procedimiento contar valores dentro de los casos (variable N_asigna).

Por último, en la tabla 8.13 se muestran las frecuencias de la variable N_asigna. Sorprende la variabilidad de las respuestas obtenidas, ya que uno de cada tres entrevistados considera que ninguna asignatura propone libros de lectura obligatoria (0 asignaturas), el 24,1% consideran que tan sólo una asignatura obliga a leer libros, y el 8,4% cree que son tres las asignaturas que proponen libros de lectura obligatoria.

Buscando fijar los conocimientos aprendidos en esta sección, recomendamos –utilizando el archivo *Vida Cotidiana*– elaborar una variable que sintetice toda la información proporcionada por la pregunta del nivel de equipamientos del hogar (pregunta 20, variables de la b8 a la b20). Se pide, en definitiva, elaborar una variable que clasifique a la población entrevistada según el número de equipamientos dis-

Número de asignaturas que proponen libros de lectura obligatoria					
		Frecuencia	**Porcentaje**	**Porcentaje válido**	**Porcentaje acumulado**
Válidos	0	65	34,0	34,0	34,0
	1	46	24,1	24,1	58,1
	2	16	8,4	8,4	66,5
	3	29	15,2	15,2	81,7
	4	13	6,8	6,8	88,5
	5	12	6,3	6,3	94,8
	6	6	3,1	3,1	97,9
	8	4	2,1	2,1	100,0
	Total	191	100,0	100,0	

Tabla 8.13. Número de asignaturas que obligan a leer libros.

ponibles en su hogar[100]; para posteriormente elaborar un *índice* que permita una aproximación al estatus social.

8. Selección de casos mediante criterios condicionales

Una de las razones de la enorme variabilidad de las respuestas mostradas en la tabla 8.12 puede ser la existencia de estudiantes de varios cursos, con diferentes asignaturas en cada curso. Como se desprende al analizar las respuestas de la variable v46, el 38,6% de los estudiantes entrevistados están en primero, un 34% pertenecen a segundo curso, mientras que el resto cursan asignaturas de tercero y cuarto.

Sería más correcto analizar el número de asignaturas que obligan a leer diferenciando el curso de los estudiantes; realizando este análisis en el primer curso y –posteriormente– en el curso segundo. Para ello utilizaremos la selección de casos mediante criterios condicionales, que se activa con el menú *Datos⇒Seleccionar Casos*[101].

100. Dicho de otro modo, se trata de crear una variable donde se muestre el número de equipamientos presentes en el hogar de cada entrevistado.
101. Conviene recordar que este procedimiento se utilizó en la secciones 6.4, pero en aquel momento se empleó con el fin de depurar la información del archivo de datos.

(v46) Curso actual					
		Frecuencia	**Porcentaje**	**Porcentaje válido**	**Porcentaje acumulado**
Válidos	1	70	36,6	36,6	36,6
	2	65	34,0	34,0	70,7
	3	48	25,1	25,1	95,8
	4	6	3,1	3,1	99,0
	No responde	2	1,0	1,0	100,0
	Total	191	100,0	100,0	

Tabla 8.14. Distribución de frecuencias de v41 (curso actual).

Situados en el cuadro de diálogo de la figura 8.20 será necesario marcar *Si se satisface la condición*, y posteriormente el botón *Si la opción...* para activar el criterio condicional elegido, lo que dará origen a la figura 8.21. En ésta se selecciona la variable correspondiente, el curso actual (v46), solicitando que seleccione únicamente el valor 1; que corresponde a los estudiantes de primero.

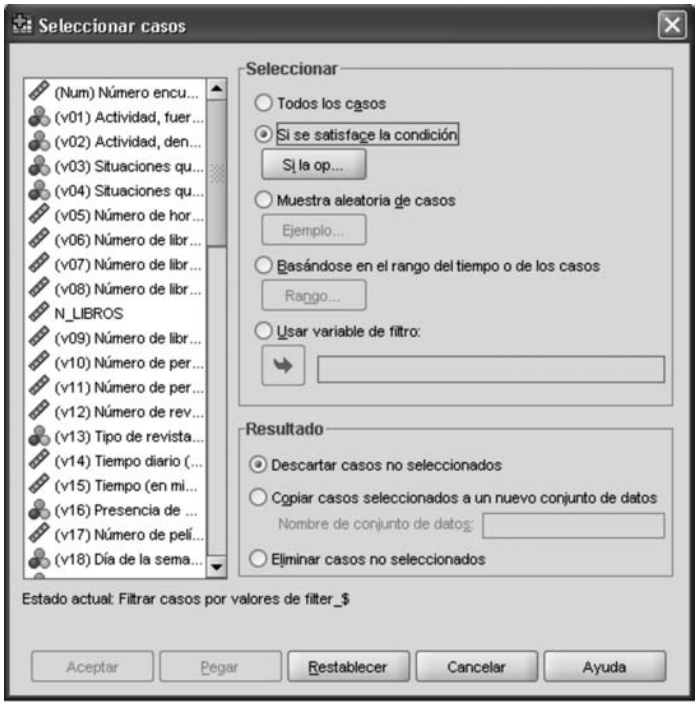


Figura 8.20. Seleccionar casos, Si se satisface la condición.

Pulsando *Continuar* en la figura 8.21 se accede al menú principal de la selección de casos (figura 8.20) y, tras comprobar que los casos no seleccionados serán *Filtrados* (opción por defecto), se pulsa el botón *Aceptar* para llevar a cabo la ejecución del procedimiento.

Figura 8.21. Selección de casos: condición lógica.

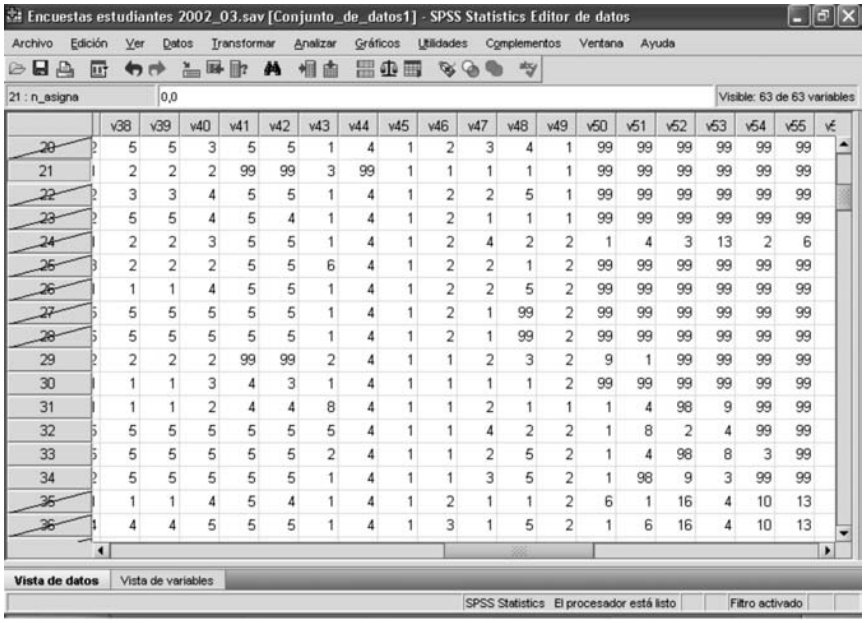


Figura 8.22. Resultado de la selección.

En la figura 8.22 se muestra el resultado de la selección, pudiendo apreciar la diferencia entre los valores seleccionados y los filtrados (estos últimos aparecen tachados en el número de caso). Obsérvese la opción *Filtro activado* en la Barra de Estado (esquina inferior derecha). A partir de este momento todos los análisis solicitados se llevarán a cabo con los casos seleccionados, esto es, con los estudiantes de primero (V46 = 1).

Solicitando, por ejemplo, las frecuencias de N_asigna, se obtendrán el *número* de asignaturas que obligan a leer en primero (tabla 8.15). Debe tenerse en cuenta que esta tabla está respondida por 70 personas (los estudiantes de primero) y no por las 191 que componen la muestra total.

Número asignaturas que proponen libros de lectura obligatoria

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	0	26	37,1	37,1	37,1
	1	24	34,3	34,3	71,4
	2	8	11,4	11,4	82,9
	3	8	11,4	11,4	94,3
	4	4	5,7	5,7	100,0
	Total	70	100,0	100,0	

Tabla 8.15. Número de asignaturas que obligan a leer libros (respuestas de los estudiantes de primero).

Las escasas diferencias entre este resultado y el obtenido en la tabla 8.13 nos lleva a negar el planteamiento con el que se originaba este apartado, cuando se señalaba que las razones de la gran variabilidad pudiera estar originado por la consideración conjunta de estudiantes de varios cursos. En cualquier caso, este planteamiento nos ha permitido explicar la selección de casos mediante criterios condicionales, objetivo último de esta exposición.

Cambiando la condición lógica de la figura 8.21, seleccionando v46=2, y solicitando las frecuencias de N_asigna se obtiene la tabla 8.16, que muestra el número de asignaturas de lectura obligatoria en segundo curso. Como puede apreciarse al observar el total de la tabla, se está trabajando con la información de 65 personas, los 65 entrevistados que están en segundo.

Número asignaturas que proponen libros de lectura obligatoria					
		Frecuencia	**Porcentaje**	**Porcentaje válido**	**Porcentaje acumulado**
Válidos	0	31	47,7	47,7	47,7
	1	8	12,3	12,3	60,0
	2	4	6,2	6,2	66,2
	3	9	13,8	13,8	80,0
	4	3	4,6	4,6	84,6
	5	4	6,2	6,2	90,8
	6	2	3,1	3,1	93,8
	8	4	6,2	6,2	100,0
	Total	65	100,0	100,0	

Tabla 8.16. Número de asignaturas que obligan a leer libros (respuestas de los estudiantes de segundo).

Si se repitiera la operación con los que cursan tercero se trabajaría con 48 personas, y con 6 si se tomaran en consideración los que están en cuarto curso. Este ejemplo muestra el principal problema que presenta este procedimiento: al *segmentar* la muestra total en subgrupos más pequeños se pierde precisión y un aumento importante del error que puede poner en duda la generalización de los resultados (fin último de la investigación con encuesta). Esto es, esas 6 personas de cuarto curso, ¿pueden considerase representativas de ese curso? Creemos que no. ¿Y los 48 entrevistados de tercero? Respondemos esta pregunta utilizando una cita de Alvira (2004: 101) cuando señala que, en universos finitos, el investigador debe indicar las dificultades de generalización de los hallazgos realizados sobre una muestra inferior a 50 casos[102].

Para terminar debe tenerse en cuenta que la selección de casos permanece activada hasta que se marca, dentro de la figura 8.20, la opción *todos los casos*. Para evitar errores recomendamos prestar atención *siempre* al total de las tablas, que nos indicará hasta que punto estamos considerando toda la muestra o tan sólo unos casos.

102. Se trata, más bien, de una recomendación del Estándar de Calidad en la Investigación de mercados (ECIM), elaborado por ANEIMO y publicado en castellano por ESOMAR (2000: 256-257).

Fijar los conocimientos aprendidos es el fin de toda actividad docente. Con este objetivo se proponen dos ejercicios a realizar con el archivo *Vida Cotidiana*. La hora de realización de las compras (pregunta 53, variables c52 y c53), ¿presenta variación según el sexo del entrevistado? ¿Y considerando el tamaño del municipio? En caso de existir diferencias importantes, indicar con detalle la *tipología de horario* que siguen los residentes en cada tamaño del municipio.

9. Segmentar archivo

La selección de casos es un recurso muy útil cuando se desea analizar los datos de una parte de la muestra; por ejemplo los hombres, las mujeres, los que estudian Sociología, los que están en primer curso, etc. Ahora bien, no es adecuado utilizar este procedimiento cuando el objetivo sea comparar una variable en varias partes de la muestra; por ejemplo comprar el número de asignaturas que obligan a leer libros en hombres y mujeres, entre los que estudian sociología y sus amigos, entre los de primero y los de segundo...

Cuando el objetivo de la investigación es comparar una variable en varios *estratos muestrales* es más adecuado utilizar el procedimiento *segmentar archivo* situado dentro del menú *Datos*. Seleccionando consecutivamente *Datos*⇒*Segmentar archivo* aparece el cuadro de diálogo que se muestra en la figura 8.23. La opción por defecto (Analizar todos los casos, no crear los grupos) indica que la segmentación se encuentra inactiva.

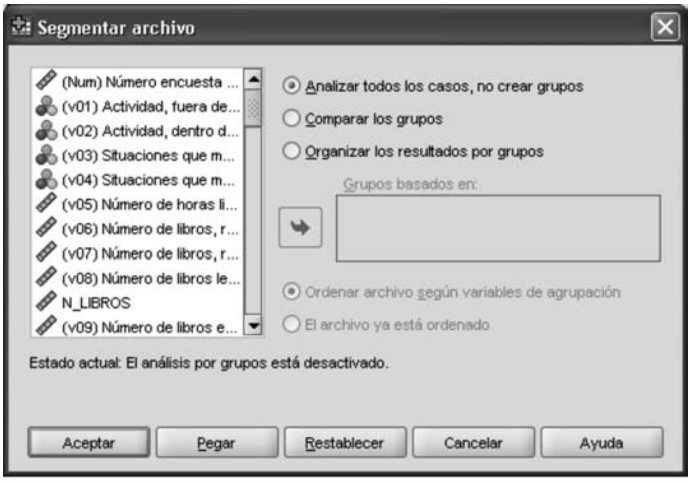


Figura 8.23. Segmentar archivo: segmentación inactiva.

Para conocer los libros de lectura obligatoria por cursos –por seguir con el ejemplo planteado en la sección anterior– basta con insertar la variable *curso actual* (v46) en la ventana situada en el centro de la pantalla, opción *Comparar los grupos* (figura 8.24), y pulsar *Aceptar*. Tras solicitar las frecuencias se obtendrán los resultados mostrados en la tabla 8.17. Antes de proceder con su interpretación, hay que tener en cuenta que este procedimiento presenta los mismos problemas que la *segmentación*: la pérdida de representatividad al analizar grupos formados por pocas unidades.

Es posible obtener los resultados en tablas distintas, marcando –en la figura 8.23– la opción *Organizar los resultados por grupos*. Por último, existe también la opción de ordenar los datos en el archivo de casos seleccionando *Ordenar archivo según variable de agrupación*. En tal caso la variable v46 aparecerá ordenada en orden ascendente, consiguiendo el mismo resultado que el obtenido cuando se utiliza el procedimiento *Ordenar casos* (expuesto en la sección 6.3)

Figura 8.24. Segmentar archivo, utilizando la v46 como criterio de segmentación.

Será necesario volver a seleccionar la opción *Analizar todos los casos, no crear los grupos* para devolver al archivo a su posición original; para des-seleccionar la segmentación.

Buscando *fijar* los conocimientos aprendidos en esta sección, antes de considerar nuevos contenidos, recomendamos –utilizando el archivo de la investigación sobre *Vida Cotidiana*– considerar si el sexo del entrevistado presenta alguna influencia en el hecho de realizar compras solo o acompañado (pregunta 54, variables c54 y c55), así como en el transporte utilizado para acudir a comprar (pregunta 55, variables c56 y c57). Y

Número asignaturas que obligan a leer libros						
(v46) Curso actual			**Frecuencia**	**Porcentaje**	**Porcentaje válido**	**Porcentaje acumulado**
1	Válidos	0	26	37,1	37,1	37,1
		1	24	34,3	34,3	71,4
		2	8	11,4	11,4	82,9
		3	8	11,4	11,4	94,3
		4	4	5,7	5,7	100,0
		Total	70	100,0	100,0	
2	Válidos	0	31	47,7	47,7	47,7
		1	8	12,3	12,3	60,0
		2	4	6,2	6,2	66,2
		3	9	13,8	13,8	80,0
		4	3	4,6	4,6	84,6
		5	4	6,2	6,2	90,8
		6	2	3,1	3,1	93,8
		8	4	6,2	6,2	100,0
		Total	65	100,0	100,0	
3	Válidos	0	6	12,5	12,5	12,5
		1	12	25,0	25,0	37,5
		2	2	4,2	4,2	41,7
		3	12	25,0	25,0	66,7
		4	6	12,5	12,5	79,2
		5	6	12,5	12,5	91,7
		6	4	8,3	8,3	100,0
		Total	48	100,0	100,0	
4	Válidos	1	2	33,3	33,3	33,3
		2	2	33,3	33,3	66,7
		5	2	33,3	33,3	100,0
		Total	6	100,0	100,0	
No responde	Válidos	0	2	100,0	100,0	100,0

Tabla 8.17. Número de asignaturas que obligan a leer libros (resultado de segmentar archivo).

el estado civil, ¿presenta alguna influencia en estas conductas? Volver a realizar los ejercicios planteados en el último párrafo de la sección 8.8, pero en esta ocasión utilizando el procedimiento *segmentar archivo* en lugar de la *selección de casos*.

10. Anexo 1: Lenguaje de sintaxis de los análisis realizados en el capítulo

En el apartado 7 del capítulo VII se ha explicado el origen de cada uno de estos mandatos (pulsando el botón Pegar en el cuadro de diálogo correspondiente), así como el proceso de *ejecución* de cada uno.

Apartado 2: Recodificación automática

```
AUTORECODE
   VARIABLES=v41 /INTO V41recod
   /PRINT.
```

Apartado 3: Categorizar variables

```
*Agrupación visual.
*v05.
RECODE v05
   ( MISSING = COPY )
   ( LO THRU 25 =1 )
   ( LO THRU 349.67 =2 )
   ( LO THRU 674.34 =3 )
   ( LO THRU HI = 4 )
   ( ELSE = SYSMIS ) INTO v05_ca_1.
VARIABLE LABELS v05_ca_1 '(v05) Número de horas libres que dispone a la'+
     'semana de ocio o diversión (Categorizada)'.
FORMAT v05_ca_1 (F5.0).
VALUE LABELS v05_ca_1
   1 '<= 25'
   2 '26 - 350'
```

```
   3 '351 - 674'
   4 '675+'.
MISSING VALUES v05_ca_1 ( ).
VARIABLE LEVEL v05_ca_1 ( ORDINAL ).
EXECUTE.

*Agrupación visual.
*v05.
RECODE v05
   ( MISSING = COPY )
   ( LO THRU 20 =1 )
   ( LO THRU 40 =2 )
   ( LO THRU 50 =3 )
   ( LO THRU HI = 4 )
   ( ELSE = SYSMIS ) INTO v02_ca_2.
VARIABLE LABELS v02_ca_2 '(v05) Número de horas libres que dispone a la'+'sema-
    na de ocio o diversión (Categorizada)'.
FORMAT v02_ca_2 (F5.0).
VALUE LABELS v02_ca_2
   1 '<= 20'
   2 '21 - 40'
   3 '41 - 50'
   4 '51+'.
MISSING VALUES v02_ca_2 ( ).
VARIABLE LEVEL v02_ca_2 ( ORDINAL ).
EXECUTE.
```

Apartado 4: Recodificar en las mismas variables

```
RECODE v43 (4=1) (7=1) (9=1) (5=3) (6=4) (8=4) (10=4).
EXECUTE.
VALUE LABELS V43 1"Con padres" 2"Con amigos en piso compartido" 3"Con pare-
    ja" 4"Otras situaciones".
FREQUENCIES
   VARIABLES=v43
   /ORDER= ANALYSIS.
```

Apartado 5: Recodificar en distintas variables

```
RECODE v43 (1=1) (2=2) (3=3) (4=1) (7=1) (9=1) (5=3) (6=4) (8=4) (10=4) INTO
    v43recod.
VARIABLE LABELS v43recod 'v43 recodificada'.
EXECUTE.
```

Apartado 6: Cálculos y operaciones

```
COMPUTE N_LIBROS=V06+V08.
FREQUENCIES
    VARIABLES=N_libros
    /ORDER= ANALYSIS.
RECODE N_libros (99 THRU 105=99).
MIS VAL N_libros (99).
FREQUENCIES
    VARIABLES=N_libros
    /ORDER= ANALYSIS.
RECODE N_libros (1 thru 2=1) (3 thru 4=2) (5 thru 7=3) (8 thru 10=4) (11 thru
    30=5).
VALUE LABELS N_libros 0"0 libros" 1"1-2 libros" 2"3-4 libros" 3"5-7 libros" 4"8,
    9 y 10 libros" 5"Más de 10".
FREQUENCIES
    VARIABLES=N_libros
    /ORDER= ANALYSIS.
```

Apartado 7: Creación de nuevas variables uniendo valores en las variables de origen

```
COUNT N_asigna = v50 v51 v52 v53 v54 v55 v56 v57 v58 (1 thru 98).
VARIABLE LABELS N_asigna 'Número asignaturas proponen libros de lectura obli-
    gatorios'.
EXECUTE.
FREQUENCIES
    VARIABLES=N_asigna
    /ORDER= ANALYSIS.
```

Apartado 8: Selección de casos mediante criterios condicionales

```
FREQUENCIES
    VARIABLES=v46
    /ORDER= ANALYSIS.

USE ALL.
COMPUTE filter_$=(v46 = 1).
VARIABLE LABEL filter_$ 'v46 = 1 (FILTER)'.
VALUE LABELS filter_$ 0 'No seleccionado' 1 'Seleccionado'.
FORMAT filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE.

FREQUENCIES
    VARIABLES=N_asigna
    /ORDER= ANALYSIS.

USE ALL.
COMPUTE filter_$=(v46 = 2).
VARIABLE LABEL filter_$ 'v46 = 2 (FILTER)'.
VALUE LABELS filter_$ 0 'No seleccionado' 1 'Seleccionado'.
FORMAT filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE.

FREQUENCIES
    VARIABLES=N_asigna
    /ORDER= ANALYSIS.

USE ALL.
EXECUTE.
```

Apartado 9: Segmentar archivo

```
SORT CASES BY v46.
SPLIT FILE
    LAYERED BY v46.
```

```
FREQUENCIES
    VARIABLES=N_asigna
    /ORDER= ANALYSIS.

FILTER OFF.
```

Capítulo IX

Tablas de contingencia de dos variables

1. Objetivos didácticos del capítulo

Toda la exposición realizada hasta el momento se ha centrado en el análisis de una variable (univariante): el séptimo capítulo se dedicó a cómo interpretar la distribución de una variable, mientras que en el octavo se explicó la creación de nuevas variables y la modificación de los valores de las variables existentes. Hasta este momento todos los análisis se han realizado variable a variable, tomando cada variable por separado. Ahora bien, en los últimos apartados del octavo capítulo (secciones 8.8 y 8.9) se ha realizado una ligera introducción al análisis bivariable al llevar a cabo el análisis de una variable considerando los diferentes valores de una segunda variable. En estos apartados se analizó el número de asignaturas que obligan a leer libros en función del curso del entrevistado, bien utilizando *criterios condicionales* (apartado 8.8) o mediante la *segmentación del archivo* (apartado 8.9), que ha servido de iniciación al análisis bivariable.

El presente capítulo está dedicado al análisis conjunto de dos variables, análisis bivariable, centrando su atención en los *cruces de tablas*, *tablas cruzadas*, o *tablas de contingencia*; una de las herramientas más utilizadas por el analista de encuestas. La utilización de tablas de contingencia aporta información conjunta de dos (o más) variables mostrando las respuestas de una en función de la otra; indicando el valor que toma la primera variable cuando la segunda tiene un determinado valor. Como tendremos ocasión de comprobar a lo largo del capítulo, esta herramienta supone grandes mejoras frente a los *criterios condicionales* o la *segmentación de archivo* presentado en el capítulo anterior.

Al igual que procedimos a lo largo de todo el trabajo, la explicación se llevará a cabo utilizando ejemplos realizados con el archivo de datos obtenido del cuestionario presentado en el segundo capítulo, sección 2.7 (ENCUESTAS ESTUDIANTES 2002_03.SAV). Los ejercicios propuestos en el capítulo nueve de los *materiales complementarios* deben realizarse con el archivo "Encuestas estudiantes (SIETE promociones). sav".

2. Elaboración de tablas de contingencia con dos variables

La elaboración de tablas de contingencia se encuentra en el menú Analizar, dentro del submenú Estadísticos descriptivos, de modo que para elaborar una tabla hay que seleccionar *Analizar⇒Estadísticos descriptivos⇒Tablas de contingencia*, lo que da paso al cuadro de diálogo de la figura 9.1 Realizaremos una tabla muy sencilla para comenzar con la explicación, teniendo presente que el objetivo es conocer las actividades de ocio fuera del hogar que caracterizan y diferencian a los hombres de las mujeres. Para elaborar esta tabla se selecciona una variable para las filas y se hace un clic en la primera de los *botones-flecha* de la figura 9.1. En este caso se ha elegido la primera variable del cuestionario (v01) que recoge la actividad, fuera de casa, que más gusta hacer cuando se dispone de tiempo libre. La variable seleccionada pasará a la ventana *Filas*.

Conviene recordar que esta variable ha sido recodificada en el capítulo, concretamente en el apartado dedicado a la *recodificación* en las mismas variables (8.4). De modo que, estrictamente hablando, no trabaremos con v01 sino con v01bis; puesto que se ha realizado una recodificación *en distintas variables*.

Figura 9.1. Cuadro de diálogo Tablas de contingencia.

Es importante tener claro el resultado del proceso para apreciar la transformación experimentada; pasando de una variable de 18 categorías de respuesta a 8. Esto se ha producido como consecuencia de elaborar un denominador común definido como "otras", que agrupa las categorías "ir al teatro" (codificada con el valor 7), "ir a conciertos" (9), "leer libros" (10), "otras" (14), "ninguna en particular" (15), "quedar con amigos" (16) y "quedar con el novio" (17). Otra de las transformaciones ha consistido en unir dos categorías debido a la similitud temática entre ellas: se trata de "ir de excursión" (codificada con el valor 4) e "ir al monte" (18), que han sido unidos en una única categoría. Es importante destacar que la opción "otras" supera ligeramente el 10% de los casos. El escaso tamaño muestral nos ha llevado a tomar esta decisión.

De esta exposición se deduce que antes de realizar un cruce de tablas es necesario que las variables a cruzar hayan sido analizadas y, cuando sea preciso, proceder a su transformación con el fin de que presenten categorías con un número aceptable de respuestas. Las variables "no agrupadas" presentan dos problemas:

- Por un lado es imposible generalizar tomando en consideración un escaso número de entrevistados. ¿Qué generalización podemos hacer de las personas que han señalado que en su tiempo libre lo que más les gusta es ir al teatro?, opción que ha sido elegida por 2 entrevistados que suponen un 1% de la muestra. Que se trate de dos mujeres, ¿quiere decir que las mujeres prefieren el teatro más que los hombres? El escaso número de elecciones impide realizar tal afirmación; no es posible extraer conclusiones de una categoría elegida únicamente por dos personas.
- En segundo lugar, y como veremos en la sección 3, los tests estadísticos utilizados para conocer la relación entre variables no funcionan correctamente cuando la tabla analizada tiene muchas *celdillas* con pocas respuestas.

Finalizada la explicación de la variable en filas, posteriormente se selecciona la variable V49, que corresponde al sexo, y se coloca en la ventana de las *columnas*[103]. Antes de proceder de esta forma es preciso conocer como se distribuye el sexo, puesto que se trata de una variable que no ha sido tratada con anterioridad. Las frecuencias de v49 desvelan que el 39% de los entrevistados son hombres y un 61% de mujeres. Pulsando el botón *Aceptar* obtenemos una tabla de contingencia, en su formato más sencillo, que se muestra en la tabla 9.1.

103. Las variables situadas en las columnas son conocidas como *variables de identificación*, *cabeceras*, o *variables cabecera*.

Resumen del procesamiento de los casos

	Casos					
	Válidos		Pérdidos		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
(v01bis) Actividad, fuera de casa, que más te gusta hacer cuando dispones de tiempo libre * (v49) Género	180	94,2%	111	5,8%	191	100,0%

Tabla de contingencia (v01bis) Actividad, fuera de casa, que más te gusta hacer cuando dispones de tiempo libre * (v49) Género

		Genero		
		Hombre	Mujer	Total
v01bis) Actividad, fuera de casa, que más te gusta hacer cuando dispones de tiempo libre * (v49) Género	Beber, ir de copas	26	20	46
	Bailar	0	12	12
	Hacer deporte	12	8	20
	Ir de excursión y al monte	6	6	12
	Viajar	12	28	40
	Ir al cine	0	10	10
	Practicar alguna afición o hobby	12	8	20
	Otras[104]	2	18	20
Total		70	110	180

Tabla 9.1. Cruce de tablas entre actividad fuera de casa que más le gusta hacer en su tiempo libre (v01bis) y sexo (v49).

Comenzaremos la interpretación de esta tabla analizando el "resumen del procesamiento de los casos" que informa sobre el número de datos analizados (191), los casos válidos (180, el 94,2%) y los perdidos (11, un 5,8%). Posteriormente procedemos con

104. Tener en cuenta que dividir más esta categoría supondrían más celdillas con menos de 5 entrevistados.

el análisis de los *marginales*, los totales de filas y columnas, para ocuparnos –más adelante– de las celdillas resultantes de la intersección de filas y columnas. A la derecha de la fila se muestran los valores totales de la variable filas (v01bis): el número 46 que aparece en la parte derecha de la primera fila corresponde al número de personas que declaran que lo que más les gusta hacer fuera de casa es beber, ir de copas; 12 entrevistados eligen bailar, 20 hacer deporte, 40 viajar... Respecto a las columnas, se trata de una muestra compuesta por 70 hombres y 110 mujeres. Esta diferencia en favor de las mujeres precisará, en el apartado 9.4.1, utilizar un determinado tipo de porcentajes que permite *mitigar* tales diferencias.

3. Utilización de test estadísticos para conocer la relación entre variables nominales

Con el fin de conocer hasta que punto las variables utilizadas en la tabla 9.1 están relacionadas se han desarrollado una serie de medidas que –en un solo índice– señalan la existencia de relación entre dos variables, así como el grado de *asociación* y su dirección. Son medidas cuyos valores oscilan entre un valor mínimo indicativo de la ausencia de asociación (normalmente el cero) y un valor máximo que indica asociación perfecta (el uno o el menos uno). Un valor superior a cero (positivo) indicará relación directa, mientras que un valor inferior a cero (negativo) muestra relación inversa[105]. García Ferrando (1985: 217-222) presenta las características que tiene que cumplir una buena *medida de asociación* entre variables: debe indicar *si existe o no una asociación significativa* entre variables, y *cuantificar la fuerza* de esa asociación. El siguiente requisito que debe cumplir una buena medida de asociación es mostrar la *dirección* de la asociación (positiva o negativa), aunque esto únicamente es posible cuando las variables se han medido a nivel ordinal o de intervalo. La cuarta característica es describir la *naturaleza* de la asociación, referida a la distribución de las magnitudes de las variables en cada una de las *celdillas* de la tabla: la comparación de los porcentajes puede mostrar una escasa diferencia en las categorías bajas de las variables, diferencia que se acentúa en las categorías medias y aún más en las altas (relación lineal), o puede tener una tendencia totalmente irregular. Por último tienen que ser medidas *estandarizadas* o *tipificadas* que permitan comparar los índices obtenidos

105. La relación será directa o inversa si las categorías de ambas variables están medidas en el mismo orden. Recordar que en la sección donde se explicaron las escalas se señaló que la codificación de las escalas ordinales (y de intervalo) debe respetar el orden serial.

en distintas tablas. A este respecto García Ferrando (1985: 222) presenta un ejemplo donde señala una supuesta investigación que detecta que la relación entre edad e "interés por la política" es de +0,52; mientras que la relación de está última con el nivel de ingresos es del +0,35. El hecho que estas medidas sean estandarizadas implica que se puedan comparar ambos índices, con el fin de por establecer que el "interés por la política" está más relacionado con la edad que con el nivel de ingresos.

Ilustraremos la explicación del párrafo anterior considerando hasta que punto una de las medidas de asociación más conocidas, el coeficiente de *correlación* lineal de Pearson[106] (la "*r*" de Pearson), cumple cada una de estas propiedades. Supondremos para ello un coeficiente de correlación entre la edad y el nivel de ingresos de –0,87. En primer lugar una medida de asociación debe indicar si existe relación entre variables. Si tenemos en cuenta que el coeficiente de correlación puede oscilar entre -1 y +1, indicando el valor central (0) la no existencia de relación; un valor de –0,87 implicará –sin duda– una relación significativa ente ambas variables[107]. Asimismo, si la máxima relación posible entre variables es 1, un valor de 0,87 indica una relación importante. Por último es posible conocer la dirección de la asociación ya que valores cer-

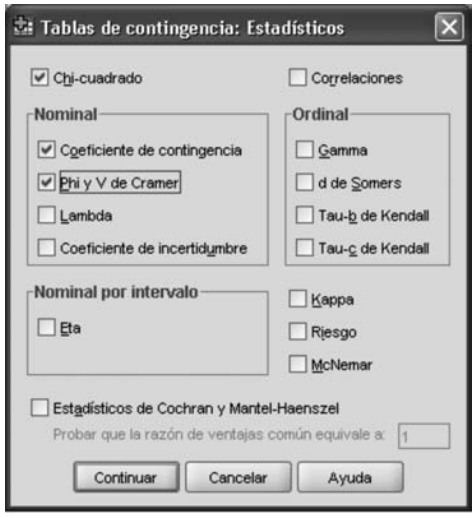


Figura 9.2. *Estadísticos* de tablas de contingencia con variables nominales.

106. En adelante nos referiremos a esta medida como *Coeficiente de Correlación*, si bien es importante indicar que en los resultados del SPSS aparece como "R de Pearson"; tal y como puede apreciarse en la tabla 9.16.

107. Para ello deberemos calcular el nivel de significación de este valor, aspecto que no expondremos aquí puesto que queda fuera de nuestros objetivos; y que puede consultarse en cualquier texto de Estadística.

canos a +1 indicarán relación directa entre variables, mientras que valores cercanos a –1 indican relación inversa, como ocurre en este ejemplo.

Este capítulo está dedicado al análisis de variables nominales y ordinales, que son las que habitualmente se utilizan en tablas de contingencia, de modo que olvidaremos las propiedades del coeficiente de correlación (variables de intervalo) para analizar si los estadísticos que miden la relación entre variables nominales y ordinales cumplen cada uno de estos criterios. A fin de llevar a cabo una exposición práctica utilizaremos el ejemplo expuesto en la tabla 9.1, si bien pulsaremos –en el cuadro de diálogo de la figura 9.1– el botón *"Estadísticos..."* para solicitar aquellos test utilizados para conocer la relación entre dos variables nominales, como son las variables sexo y actividad fuera de casa que más le gusta hacer en su tiempo libre. Tras solicitar *Chi-Cuadrado*, *Coeficiente de Contingencia*, *Phi,* y *V de Cramer* (figura 9.2) obtenemos los resultados mostrados en la tabla 9.2.

Pruebas de chi-cuadrado

	Valor	Gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	36,496(a)	7	,000
Razón de verosimilitud	45,236	7	,000
Asociación lineal por lineal	2,781	1	,095
N de casos válidos	180		

a 3 casillas (18,8%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 3,89.

Medidas simétricas

		Valor	Sig. aproximada
Nominal por nominal	Phi	,450	,000
	V de Cramer	,450	,000
	Coeficiente de contingencia	,411	,000
N de casos válidos		180	180

a Asumiendo la hipótesis alternativa.
b Empleando el error típico asintótico basado en la hipótesis nula.

Tabla 9.2. *Estadísticos* para variables nominales: Chi-Cuadrado, Phi, V de Cramer y C de Contingencia.

3.1. Relación entre variables nominales utilizando el Chi-Cuadrado

El primero de los estadísticos solicitados es el Chi-Cuadrado, que aparece en la tabla 9.2 con el nombre de Chi-Cuadrado de Pearson. Este estadístico es un contraste que tiene en cuenta la totalidad de la tabla y se emplea para saber si la relación entre estas dos variables es significativa. En el cuadro 9.1 se muestra que el Chi-Cuadrado se calcula restando en cada *celdilla* las frecuencias observadas menos las esperadas (o teóricas), multiplicando esta diferencia al cuadrado, y dividiéndola entre las frecuencias esperadas. Las frecuencias esperadas son las que hubiera tenido la tabla de no existir relación entre variables (Calvo, 1990: 146), y se obtiene de multiplicar el total de fila por el total de columna, y dividiendo el resultado entre el número de casos. La frecuencia esperada de la *celdilla* superior izquierda de la tabla 9.1, por ejemplo, se obtiene de la multiplicación (70 * 46) / 180. Las frecuencias esperadas de las dos variables utilizadas se presentan en la tabla 9.3.

Fórmula:

$$X^2 = \Sigma \frac{(FO - FE)^2}{FE}$$

FO: Frecuencias observadas
FE: Frecuencias esperadas (o teóricas)

Cálculo con los datos de la tabla 9.3:

$$\frac{(26 - 17{,}9)^2}{17{,}9} + \frac{(0 - 4{,}7)^2}{4{,}7} + \frac{(12 - 7{,}8)^2}{7{,}8} + \frac{(12 - 15{,}6)^2}{15{,}6} + \frac{(6 - 4{,}7)^2}{4{,}7} +$$

$$\frac{(12 - 15{,}6)^2}{15{,}6} + \frac{(0 - 3{,}9)^2}{3{,}9} + \frac{(12 - 7{,}8)^2}{7{,}8} + \frac{(2 - 7{,}8)^2}{7{,}8} + \frac{(20 - 28{,}1)^2}{28{,}1} +$$

$$\frac{(12 - 7{,}3)^2}{7{,}3} + \frac{(8 - 12{,}2)^2}{12{,}2} + \frac{(6 - 7{,}3)^2}{7{,}3} + \frac{(28 - 24{,}4)^2}{24{,}4} + \frac{(10 - 6{,}1)^2}{6{,}1} +$$

$$\frac{(8 - 12{,}2)^2}{12{,}2} + \frac{(18 - 12{,}2)^2}{12{,}2} = 36{,}496$$

Grados de Libertad = (f - 1) (c - 1) = (8 - 1) (2 - 1) = 7

Cuadro 9.1. Cálculo del Chi-Cuadrado.

El sumatorio de las diferencias entre las frecuencias esperadas y las observadas, multiplicadas al cuadrado y dividiéndola entre las frecuencias teóricas será, en este caso 36,496 (cuadro 9.1). Se ha señalado en el párrafo anterior que las frecuencias esperadas son las que hubiera tenido la tabla de no existir relación entre variables, de modo que si a las frecuencias obtenidas se les resta las esperadas, una gran diferencia estará indicando que existe relación entre variables. Como el estadístico Chi-Cuadrado se calcula sumando los valores de estas diferencias (al cuadrado divididas entre la frecuencia esperada), un elevado valor del Chi-Cuadrado indicará importantes diferencias entre las frecuencias observadas y las esperadas, o dicho de otro modo, existencia de relación entre variables.

Tabla de contingencia (v01bis) Actividad, fuera de casa, que más te gusta hacer cuando dispones de tiempo libre * (v49) Género

			(v49) Genero		
			Hombre	**Mujer**	**Total**
(v01bis) Actividad, fuera de casa, que más te gusta hacer cuando dispones de tiempo libre * (v49) Género	Beber, ir de copas	Recuento	26	20	46
		Frecuencia esperada	17,9	28,1	46,0
	Bailar	Recuento	0	12	12
		Frec. esperada	4,7	7,3	12,0
	Hacer deporte	Recuento	12	8	20
		Frec. esperada	7,8	12,2	20,0
	Ir de excursión y al monte	Recuento	6	6	12
		Frec. esperada	4,7	7,3	12,0
	Viajar	Recuento	12	28	40
		Frec. esperada	15,6	24,4	40,0
	Ir al cine	Recuento	0	10	10
		Frec. esperada	3,9	6,1	10,0
	Practicar alguna afición o hobby	Recuento	12	8	20
		Frec. esperada	7,8	12,2	20,0
	Otras	Recuento	2	18	20
		Frec. esperada	7,8	12,2	20,0
Total		Recuento	70	110	180
		Frec. esperada	70,0	110,0	180,0

Tabla 9.3. Tabla de contingencia con frecuencias observadas (recuento) y frecuencia esperada.

Explicaremos detenidamente lo expuesto en el párrafo anterior con ayuda de dos ejemplos ficticios mostrados en la tabla 9.4; y cuyo objetivo es analizar si existe relación entre el sexo y el tipo de ocio. Es preciso volver insistir que se trata de un ejemplo ficticio, que supone una importante simplificación de la realidad en la que existen únicamente dos tipos de ocio (beber y bailar), pero que consideramos puede resultar muy útil para el tema que nos ocupa (explicar la existencia –o ausencia– de relación entre variables). Observando la tabla de la izquierda se aprecia con claridad la ausencia de relación entre el sexo de los entrevistados y el tipo de ocio: 25 mujeres dedican su ocio a beber, y otras 25 a bailar. Lo mismo ocurre con los hombres: 25 lo emplean su tiempo de ocio en beber, y otros 25 en bailar.

La tabla de la derecha, sin embargo, presenta grandes diferencias en las pautas de ocio de los hombres y las mujeres: 49 hombres (de una muestra de 100) muestran su preferencia por beber, y 48 mujeres se decantan por bailar. Tan sólo 1 hombre emplean su tiempo libre bailando, y 2 mujeres bebiendo. La tabla central, rotulada con el nombre B, muestra una ligera influencia, influencia que será necesario ver hasta que punto es importante (significativa) o no. Considerando estos ejemplos, ¿tiene claro el lector que tabla C está desvelando una asociación entre variables, algo que no ocurre en la tabla A? La tabla A, así concebida, sería similar a la tabla de frecuencias esperadas –tabla de no relación entre variables– de modo que cuanto más diferentes sean los datos obtenidos respecto a esta tabla, mayor será la relación entre variables. Esto es lo que intentamos explicar dos párrafos más arriba.

	Sexo	
	Hombre	**Mujer**
Beber	25	25
Bailar	25	25

Tabla A

	Sexo	
	Hombre	**Mujer**
Beber	30	20
Bailar	20	20

Tabla B

	Sexo	
	Hombre	**Mujer**
Beber	49	2
Bailar	1	48

Tabla C

Tabla 9.4. Tabla de contingencia (ejemplo ficticio).

Dijimos más arriba que un gran valor del Chi-Cuadrado indicará relación entre variables, pero ¿a partir de que límite definimos el *gran valor* del Chi-Cuadrado? Para responder a esta pregunta es preciso considerar la columna *sig. asintótica (bilateral)* o *sig. aproximada* de los estadísticos mostrados en la tabla 9.2. Estos valores están indicando la *probabilidad de equivocarnos* al señalar que existe relación entre variables, probabilidad expresada en tantos por uno. Considerando la significación del

Chi–Cuadrado, por ejemplo, la probabilidad de equivocarnos si decimos que existe relación entre *sexo* y *actividad fuera de casa que más gusta hacer cuando se dispone de tiempo libre* es de 0,000, o lo que es lo mismo del 0,0%. Es decir, al tratarse de una probabilidad de equivocarnos prácticamente nula decimos que hay relación entre variables, que los valores de la variable "actividad fuera de casa..." presenta variación según el sexo del entrevistado.

Alguien se preguntará por el límite de este valor, ¿cuando se considera que una *probabilidad de equivocarnos* es lo suficientemente alta para que no podamos decir que exista relación entre variables? En el ámbito de la investigación con encuestas se recomienda no considerar valores superiores al 0,05, es decir, probabilidades mayores del 5% son elevadas para afirmar que existe relación entre variables (Avira, 2000: 109). Si algún lector tiene problemas para recordar estos valores recomendamos que piense en el término *nivel de confianza* (que indica la probabilidad de acertar) y recuerde los niveles de confianza manejados en otras asignaturas: es muy probable que haya manejado niveles de confianza del 90% y del 95%. Un *nivel de confianza* del 95% indica una probabilidad de acertar del 95%, esto es, un nivel de *significación* (probabilidad de equivocarnos) del 5%, o dicho en tantos por uno, probabilidad de acertar del 0,95 y de equivocarnos del 0,05.

Veamos, por un momento, el valor del Chi-Cuadrado y la *significación* de la tabla 9.5. La *significación* 0,085 está indicando una probabilidad de equivocarnos del 8,5%, una probabilidad muy superior del 5% (0,05) que recomendamos en el párrafo anterior, lo que implica la ausencia de relación entre *actividad fuera de casa...* y *titulación* (sociología y resto); que son las dos variables que forman esta tabla.

3.2. Consideraciones a tener en cuenta en la utilización del Chi-Cuadrado

Para utilizar correctamente el Chi-Cuadrado los datos deben cumplir una serie de requisitos. Expondremos aquí los señalados por Reynolds (1984: 19): el primero postula que la muestra sea *aleatoria simple,* aspecto que se cumple en contadas ocasiones debido a que la selección de los entrevistados casi nunca se realiza de forma totalmente aleatoria, puesto que los sistemas de rutas y cuotas utilizados para localizar a los individuos elimina la aleatoriedad muestral. Respecto al otro término señalado en cursiva, muy pocas veces se utilizan muestras *simples* en la investigación con encuesta al recurrir normalmente a muestreos estratificados. Como el cumplimiento de ambos supuestos se realiza en contadas ocasiones, el valor del Chi-Cuadrado es el que tendrían nuestros datos si hubiéramos cumplido los citados requisitos, de modo que cuando no se cumplan consideraremos este valor a modo indicativo.

Tabla de contingencia (v01bis) Actividad, fuera de casa, que más te gusta hacer cuando dispones de tiempo libre * (Titulac) Titulación (Sociología/no sociología)[108]

		Titulación (Sociología/no sociología)		
		Sociología	No sociología	Total
(v01bis) Actividad, fuera de casa, que más te gusta hacer cuando dispones de tiempo libre	Beber, ir de copas	16	30	46
	Bailar	8	4	12
	Hacer deporte	12	8	20
	Viajar	26	14	40
	Ir al cine	6	4	10
	Practicar alguna afición o hobby	12	8	20
	Otras	20	12	32
Total		100	80	180

Pruebas de chi-cuadrado

	Valor	Gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	11,109(a)	6	,085
Razón de verosimilitud	11,153	6	,084
Asociación lineal por lineal	3,160	1	,075
N de casos válidos	180		

a 1 casillas (7,1%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 4,44.

Tabla 9.5. Ejemplo de tabla de contingencia con relación no significativa.

Otro requisito citado por Reynolds es que las categorías de las variables sean exhaustivas y mutuamente excluyentes. El tercer requisito recomienda no considerar el valor del Chi-Cuadrado cuando existan en la tabla muchas celdillas (un 20%) con frecuencias esperadas menores que 5, puesto que en esta situación no se cumple

108. Esta variable fue creada en el ejercicio 12 de la sección 8.10, y diferencia a los estudiantes de sociología del resto.

uno de los supuestos fundamentales de la distribución Chi-Cuadrado. Una de las formas para evitar esta situación es no utilizar el Chi-Cuadrado con pequeños tamaños muestrales. Otra de las estrategias para solucionar este problema es recodificar las variables con muchas categorías, uniendo las celdillas que tienen pocas respuestas con otras categorías similares. Como hemos señalado en el capítulo anterior el criterio utilizado para recodificar es que las categorías unificadas tengan una *significación* temática, eliminando así los errores muestrales altos que tienen las categorías con pocos sujetos. No obstante, en tablas de 2 x 2 no es posible realizar recodificaciones cuando alguna de las celdillas tiene una frecuencia esperada menor que 5, de modo que la única solución es utilizar el Test Exacto de Fisher en vez del Chi-Cuadrado, que el SPSS muestra automáticamente en el momento que se dan estas condiciones.

En la tabla 9.6 se muestra un ejemplo de un cruce de tablas *inaceptable*, un cruce de tablas donde no es posible interpretar el Chi-Cuadrado puesto que el 62% de las celdillas tienen una frecuencia esperada menor que cinco. Una segunda razón –tan importante como la anterior– es el escaso número de entrevistados que eligen algunas categorías (ir al teatro, leer libros), etc.[109]

La siguiente consideración a tener en cuenta en la utilización de este estadístico está relacionada con el hecho que el Chi-cuadrado utiliza una distribución de probabilidad continua como una *aproximación* a una distribución discreta. Esta *aproximación* indica que existe una relativa incorrección en el cálculo del Chi-Cuadrado, incorrección que es mayor a medida que disminuye el número de categorías. Esta incorrección es prácticamente nula en variables discretas con múltiples categorías, pero alcanza valores importantes en variables dicotómicas. Por ello para tablas de 2 x 2 Yates propuso la *Corrección de Continuidad* que lleva su nombre, calculada restando 0,5 al resultado FO - FT del numerador en la fórmula del Chi-cuadrado (cuadro 9.2). Como el objetivo es restar 0,5, si el resultado FO - FT es negativo será necesario sumar 0,5. El programa SPSS calcula automáticamente la corrección de Yates en tablas de contingencia de 2 filas y 2 columnas, de modo que será necesario desviar la atención del valor Chi-Cuadrado a la Corrección de Continuidad de Yates siempre que ésta aparezca.

$$X^2 = \Sigma \frac{([FO - FT] - ó + 0{,}5)^2}{FT}$$

Cuadro 9.2. Corrección de Continuidad de Yates.

109. Obsérvese que no se trata de v01bis, sino de v01. Recordemos que v01 es la variable sin recodificar, tal y como ha sido recogida en el cuestionario.

Tabla de contingencia (v01) Actividad, fuera de casa, que más te gusta hacer cuando dispones de tiempo libre * (v49) Género

		(v49) Genero		
		Hombre	**Mujer**	**Total**
v01) Actividad, fuera de casa, que más te gusta hacer cuando dispones de tiempo libre	Beber, ir de copas	26	20	46
	Bailar	0	12	12
	Hacer deporte	12	8	20
	Ir de excursión	2	6	8
	Viajar	12	28	40
	Ir al cine	0	10	10
	Ir al teatro	0	2	2
	Ir a conciertos	0	4	4
	Leer libros	0	2	2
	Practicar alguna afición o hobby	12	8	20
	Otras	0	6	6
	Ninguna en particular	2	0	2
	Quedar con amigos/as	0	2	2
	Quedar con el novio/a	0	2	2
	Ir al monte	4	0	4
	Más de una respuesta	1	10	11
Total		71	120	191

Pruebas de chi-cuadrado

	Valor	**Gl**	**Sig. asintótica (bilateral)**
Chi-cuadrado de Pearson	55,209(a)	15	,000
Razón de verosimilitud	70,676	15	,000
Asociación lineal por lineal	3,808	1	,051
N de casos válidos	191		

a 20 casillas (62,5%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es, 74.

Tabla 9.6. Tabla de contingencia con un valor de Chi–Cuadrado erróneo; con un valor que no debe interpretarse.

En la tabla 9.7 no supone ninguna diferencia considerar el Chi-Cuadrado o la Corrección de Continuidad de Yates, pero en numerosas ocasiones nos encontramos con tablas de 2 x 2 donde el Chi-Cuadrado es significativo y no así la Corrección de Continuidad. En estos casos concluiremos que no existe relación significativa entre las variables de la tabla, puesto que en tablas de 2 x 2 la atención debe desviarse del Chi-Cuadrado a la Corrección de continuidad.

Tabla de contingencia (v28) Dispositivos en el ordenador: lectora DVD * (v49) Género

		(v49) Genero		
		Hombre	**Mujer**	**Total**
(v28) Dispositivos en el ordenador: LECTORA DVD	No/no responde	12	66	78
	Si	56	48	104
Total		68	114	182

Pruebas de chi-cuadrado

	Valor	**Gl**	**Sig. asintótica (bilateral)**	**Sig. exacta (bilateral)**	**Sig. exacta (bilateral)**
Chi-cuadrado de Pearson	28,173(b)	1	,000		
Corrección por continuidad(a)	26,554	1	,000		
Razón de verosimilitud	30,019	1	,000		
Estadístico exacto de Fisher				,000	,000
Asociación lineal por lineal	28,019	1	,000		
N de casos válidos	182				

a Calculado sólo para una tabla de 2x2.
b 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 29,14.

Tabla 9.7. Tabla de contingencia 2 x 2, Corrección de Continuidad de Yates.

Finalizaremos la exposición sobre el Chi-Cuadrado volviendo a recordar los requisitos que debe cumplir una *buena* medida de asociación. En páginas anteriores se ha utilizado el coeficiente de correlación como ejemplo para ilustrar algunas explicaciones, debido a su amplia difusión y conocimiento. Desearíamos utilizar de nuevo su

popularidad para analizar las diferencias entre éste y el Chi-Cuadrado, tal y como se presenta en el cuadro 9.3. Las dos primeras características del cuadro 9.3 se refieren específicamente a los requisitos apuntados. Ambos señalan si existe o no asociación, si bien el Chi-Cuadrado no indica el sentido de la asociación entre variables porque las variables nominales no llevan implícitas ninguna relación de orden entre sus categorías.

Chi-Cuadrado

- Afirma si existe o no asociación
- No indica el sentido de la asociación
- Sirve para variables nominales, ordinales y de intervalo.
- No exige "distribución especial" de las variables.
- No exige función especial entre ambas variables.

Coeficiente de Correlación

- Afirma si existe o no relación.
- Indica el grado de relación.
- Indica el sentido de la asociación
- Sólo sirve para variables de intervalo.
- Exige que ambas variables sigan la curva normal.
- Exige función rectilínea lineal entre las variables.

Cuadro 9.3. Diferencias entre el Chi-Cuadrado y el coeficiente de correlación (Calvo 1990: 145).

Volviendo de nuevo a la primera característica del cuadro 9.3, el Chi-Cuadrado permite afirmar si existe o no asociación (significativa) entre variables, pero no indica el grado de esta asociación. Su propia formulación, una resta al cuadrado entre frecuencias observadas y teóricas, genera que no tenga un límite superior fijo como el coeficiente de correlación. El Chi-Cuadrado es siempre positivo y puede llegar a un valor máximo de N(K-1), donde N es el tamaño de la muestra y K es el número más pequeño de filas o columnas, que en el caso de la tabla 9.1 esto implica 180*(2-1). Si el Chi-Cuadrado de la tabla 9.1 puede alcanzar un valor máximo de 180, ¿cuanta relación entre variables indicará el valor obtenido, el 36,496? Es difícil responder a esta pregunta con la información proporcionada por el Chi-Cuadrado, mucho más cuando la sensibilidad del Chi-Cuadrado al tamaño de la muestra[110] genera que dos tablas con idéntica distribución de porcentajes –pero con distinto número de casos–

110. El valor del Chi-Cuadrado varía en función del tamaño de la muestra.

presentan dos valores diferentes. Volviendo a nuestro ejemplo, una distribución porcentual idéntica a la tabla 9.8, pero basada en 360 casos, proporciona un Chi-Cuadrado de 79,992; el doble que el mostrado en la tabla de 180 casos. Estos motivos han llevado a desarrollar distintas medidas de asociación que serán presentadas en el siguiente apartado.

3.3. Estadísticos basados en el Chi-Cuadrado

Las razones apuntadas en los últimos párrafos del apartado anterior, así como la dificultad para cuantificar la relación entre variables una vez que ya se sabe que ésta es significativa, ha generado la necesidad de utilizar medidas de asociación que permitan solucionar estos problemas. Podemos definir *medida de asociación* como un índice numérico que indica la existencia, grado y dirección de la asociación entre dos variables. Las más utilizadas, con su formulación correspondiente, se presentan en el cuadro 9.4. Algunas de ellas únicamente son aplicables en tablas cuadradas, por lo que será necesario conocer en que momento debemos utilizar cada una, algo que se detalla a continuación.

La primera de estas medidas, conocida como *contingencia cuadrática media* o *Phi-Cuadrado* y calculada como se expone en el cuadro 9.4, oscila entre 0 y 1 y su magnitud indica el grado de asociación entre las variables (García Ferrando, 1985: 224). Así 0 indica ausencia de relación y 1 máxima relación entre variables[111]. Únicamente puede ser utilizado en tablas de 2 x 2 porque en tablas con más de dos categorías su valor máximo puede superar la unidad (García Ferrando, 1985: 224). Además, se trata de una medida muy sensible a la presencia de totales marginales desequilibrados (Ruiz Maya et al, 1990: 264).

En tablas que no sean de 2 x 2 el coeficiente de *Contingencia* permite solucionar algunas de las limitaciones de *Phi*, pero presenta el problema que no llega a la unidad aunque las variables estén perfectamente relacionadas[112], ya que el denominador es siempre superior al numerados (ver cuadro 9.4). En tablas cuadradas (cuando el número de filas es igual al de columnas) se utiliza el coeficiente de Contingencia dividido entre el *C máximo* (Calvo, 1990: 157-8 y Reynolds, 1984: 47). El *C máximo* es la raíz cuadrada del número de filas (o columnas) menos uno dividido entre el número de

111. En la tabla 9.2 se muestra su magnitud en el ejemplo utilizado en este capítulo. En la figura 9.2 se muestra al cuadro de diálogo donde se solicita el cálculo de tales estadísticos.

112. A los interesados en ampliar conocimientos sobre el tema recomendamos la lectura del trabajo de A. Camarero Rioja (2002), en especial las páginas 381-384 donde se presenta el origen y los problemas asociados a cada una de estas medidas.

A) *Phi:* (tablas de 2 * 2)

Datos tabla 9.1:

$$\varphi = \sqrt{\frac{X^2}{N}} = \sqrt{\frac{31{,}447}{180}} = 0{,}418$$

B) *Coeficiente de Contingencia o Coef. C de Pearson*:

Datos tabla 9.1:

$$C = \sqrt{\frac{X^2}{X^2 + N}} = \sqrt{\frac{31{,}447}{31{,}447 + 180}} = 0{,}3856$$

C) *V de Cramer*:

Datos tabla 9.1:

$$V = \sqrt{\frac{X^2}{N^* \text{ mínimo } (f-1) \text{ o } (c-1)}} = \sqrt{\frac{31{,}447}{180^*}} = 0{,}418$$

Cuadro 9.4. Estadísticos basados en el Chi-Cuadrado.

filas (o columnas[113]), de modo que en tablas de 2 x 2 este valor es 0,707 (√ [K-1/k]; √ [2-1/2]), en tablas de 3 x 3 llega a 0,81 (√ [K-1/K]; √ [3-1/3]), en tablas 4 x 4 de 0,87, y en tablas 5 x 5 el C máximo llega a 0,89 (García Ferrando, 1985: 225). El ratio C / $C_{máx}$ se interpreta de modo similar al coeficiente de correlación al cuadrado, esto es, el tanto por uno que C es de $C_{máx}$ (Calvo, 1990: 158). Esta medida, por su parte, presenta el problema que no es posible comparar tablas de diferentes tamaños.

Para solucionar el problema de la tablas rectangulares Cramer desarrollo el estadístico "V" que oscila entre 0 y 1, con independencia del tamaño de la tabla. Como puede apreciarse, el valor del *V de Cramer* en tablas cuadradas de 2 x 2 es el mismo que el obtenido por la *Phi*, como sucede en el cuadro 9.4 y en la tabla 9.2.

Veamos, mediante un ejemplo la mejora que suponen estas medidas respecto a la utilización del Chi–cuadrado, considerando para ello las tablas mostradas dentro de la tabla 9.4. La tabla A presenta un Chi–Cuadrado de 0,000, que indica la inexistencia de relación entre variables, por lo que no procede calcular el valor *Phi*. La tabla B tiene un Chi-Cuadrado de 4, un valor Phi de 0,2, con una significación de 0,046;

113. Es indiferente considerar el número de filas o de columnas puesto que se trata de una tabla cuadrada, una tabla que presenta igual número de filas que de columnas.

que indica una relación significativa *justa*, casi en el límite. La tabla "C", por su parte, presenta un Chi-Cuadrado de 88,395, con un valor Phi de 0,940 que es significativo al 0,000[114]. Un valor de Phi cercano a la unidad (o casi uno) indica que existe una gran relación entre sexo y tipo de ocio; esto es, que el beber o el bailar está muy relacionado con el sexo del entrevistado. Además, esto implica que el sexo es *determinante* en el tipo de ocio, que no influye ninguna otra variable. Estaremos de acuerdo que este valor de 0,94 es fruto de un ejemplo, que en el *mundo real* en difícil encontrar una relación de tal magnitud, una relación tan determinante. Por este motivo recomendamos –más que valorar el valor del coeficiente evaluando cuanto se aproxima a 0 o a 1– considerar la magnitud de un coeficiente en relación con el promedio; esto es, valorar si un coeficiente es el mayor o menor de todos los considerados.

Pese a las mejoras que suponen la utilización de estas medidas frente al uso del Chi-Cuadrado, Reynolds (1984: 34-49) señala los problemas que surgen a la hora de interpretar estos coeficientes: estas magnitudes se interpretan según su proximidad a 1 o 0, de modo que si están cercanas a uno la relación será importante, mientras que ésta será despreciable si están cercanas a 0. Esta es toda la información que suministran estos coeficientes, no siendo posible interpretarlos como la variación porcentual de una variable que es explicada por otra, ni reducción del error al predecir una variable mediante el conocimiento de la otra[115]. En palabras de este autor, carecen de una interpretación intuitiva: ¿Cómo interpretamos un valor de 0,29? Parece una relación débil pero no hay una medida que ayude a decidir sobre la debilidad de esta relación (Reynolds 1984: 49). Por otro lado, estas medidas se han desarrollado para solucionar algunas de las limitaciones del Chi-Cuadrado, y como éste todas son simétricas, no distinguen entre variables dependientes e independientes.

¿Y si el Chi–Cuadrado no muestra relación entre variables?, como ocurre en el ejemplo de la tabla 9.5. En este caso no se procede al cálculo de estos estadísticos. Una vez constatado que el Chi-Cuadrado no es significativo, el análisis de la tabla termina concluyendo que no existe relación entre *las actividades que más te gusta hacer (fuera de casa) cuando dispones de tiempo libre* y el hecho de *estudiar sociología u otra* carrera.

Buscando *fijar* los conocimientos aprendidos en esta sección, antes de considerar nuevos contenidos, proponemos varios ejercicios utilizando la investigación sobre *Vida Cotidiana*. El primero plantea si existe diferencia en el grado de felicidad (a54) con-

114. Debe tenerse en cuenta que en tablas de 2*2 no se interpreta el Chi-Cuadrado, sino la corrección de continuidad propuesta por Yates (cuadro 9.2 y tabla 9.7); que en estos casos es de 3,20 (significación 0,072) y de 84,67 (significación 0,000). Lo que indica que no existe relación significativa entre variables en la *tabla B*, y sí en la *tabla C*.

115. Volviendo al ejemplo anterior del coeficiente de correlación, el coeficiente de 0,8 antes apuntado está indicando que una variable explica un 64% (8 * 8 = 64) de la varianza de la otra. Esta interpretación no es posible con los coeficientes basados en el Chi-Cuadrado.

siderando el sexo (e9) de los entrevistados, esto es, si es posible afirmar que los hombres son más (o menos) felices que las mujeres. ¿Y respecto al estado civil (e12)?; ¿existe diferencia en el grado de felicidad considerando el estado civil de los entrevistados?

El tercer ejercicio propone considerar hasta que punto está relacionada la variable "clase social de pertenencia" (e26) con el nivel de equipamiento del hogar considerando conjuntamente todos los equipamientos. Nos referimos, concretamente, a la variable creada al final del apartado 8.7 utilizando la pregunta 20, que clasifica a la población entrevistada según el número de equipamientos presentes en su hogar.

4. Análisis del interior de la tabla

Una vez comprobado que existe relación entre variables se procede con el estudio del interior de la tabla con el fin de interpretar en que consiste esa relación. Se trata, siguiendo el ejemplo empleado a lo largo de este apartado, de señalar cuáles son las actividades de ocio que caracterizan a los hombres y a las mujeres. Para ello presentamos dos estrategias: cálculo e interpretación de porcentajes, e interpretación de la tabla utilizando *residuo*s.

4.1. Cálculo y diferencia de porcentajes

Volvamos de nuevo a la tabla 9.1 con el fin de analizar una de las formas más populares de interpretación de los valores de las tablas de contingencia. Basados en la idea que la interpretación de los números absolutos de las celdillas es complicada, la mayor parte de las veces se recurre al estudio de porcentajes, dejando en un lugar secundario el análisis de los números absolutos. Comenzando, por ejemplo, con el análisis de los hombres que emplean su tiempo de ocio en beber, aparecen dificultades al comparar las frecuencias absolutas de los 26 hombres que manifiestan este comportamiento (celdilla superior izquierda), con las 20 mujeres que también lo declaran (celdilla superior derecha); en la medida que cada columna tiene un distinto número de personas (70 y 110 respectivamente). Por este motivo no estamos interesados en el número absoluto puesto que para poder comparar ambas celdillas es necesario *ponderar* el número de respuestas de cada celdilla respecto al número de respuestas de esta categoría en la variable columna. Así las 26 entrevistados que declaran emplear su tiempo de ocio en beber son un 37,1% de todos los hombres entrevistados, mientras que las 20 mujeres que eligen esta misma opción son un 18,2% del total de mujeres.

Para solicitar los porcentajes en el programa SPSS será necesario, dentro del Cuadro de diálogo *Tablas de contingencia* mostrado en la figura 9.1, pulsar el recuadro *Casillas...* para obtener el cuadro de diálogo expuesto en la figura 9.3. Centraremos nuestra atención en la parte inferior izquierda titulada *Porcentajes*, que permite obtener tres tipos de porcentajes haciendo clic en el lugar correspondiente:

Figura 9.3. Botón *Casillas* dentro de cuadro de diálogo de la Tablas de contingencia. Mostrar en las casillas.

- Porcentajes de columna: calculados dividiendo la frecuencia absoluta de la celdilla entre el número de respuestas de cada una de las categorías de la variable columna (v049); tal y como se presenta en el primer ejemplo de la tabla 9.8. La variable situada en la columna se considera como independiente, y la colocada en la fila como dependiente.
  Cálculo de los porcentajes de la segunda columna: 20/110 = 0,182; 12/110 = 0,109; 8/110 = 0,073...
  Estos porcentajes se interpretan de la siguiente forma: el 18,2% de las mujeres señalan que la actividad que más les gusta hacer cuando disponen de tiempo libre es beber, porcentaje que se reduce al 10,9% cuando se trata de bailar, y al 7,3% respecto a hacer deporte.
  Si observamos la columna de los hombres se aprecia que el 37,1% elige beber cuando dispone de tiempo libre, y un 17,1% muestra su preferencia por hacer deporte. Es importante indicar que bailar no es elegida por ningún hombre.

Hasta ahora únicamente se han interpretado los porcentajes que componen cada columna, pero el análisis de tablas de contingencia se ve notablemente enriquecido al poder comparar *transversalmente* estos porcentajes, realizando comparaciones entre cada una de las celdas de las distintas columnas. Para esto es fundamental la observación de la columna total, que no es otra cosa que la frecuencia de la variable v01bis (el *marginal*). Esta columna es vital para la interpretación de tablas de contingencia puesto que permite localizar los porcentajes de cada categoría (hombres y mujeres) que se encuentran por encima y por debajo del valor marginal[116]. Así es posible apreciar que los hombres destacan (respecto del total) por sus mayores elecciones de beber (37,1% hombres y 25,6% total), hacer deporte (17,1% y 11,1%) y practicar alguna afición o hobby (17,1% y 11,1%); mientras que las mujeres eligen (más que el promedio) bailar (10,9% y 6,7%), ir al cine (9,1% y 5,6%) y otras[117] (16,4% y 11,1%). A la hora de señalar la preferencia por un tipo y otro de ocio únicamente deben considerarse las diferencias *importantes* de porcentajes. Cea D'Ancona (2004: 407), por ejemplo, considera que únicamente las diferencias superiores a cinco puntos pueden considerarse relevantes[118].

- Porcentajes de fila. Calculados con el mismo criterio pero considerando el porcentaje respecto a la variable fila (v01bis). En este caso La variable situada en la fila se considera como independiente, y la colocada en la columna como dependiente.

  Cálculo de la primera fila: 26/46 = 0,5652; 20/46 = 0, 4348.

  El segundo ejemplo de la tabla 9.8 muestra esta situación. La interpretación de estos datos se realiza de modo similar, pero teniendo que cuenta que la fila es el total. De aquellos que lo que más les gusta hacer en su tiempo libre es beber, el 56,5% son hombres y un 43,5% mujeres. De los que eligen hacer deporte, el 60% son hombres y un 40% mujeres. Del total de la muestra un 39% (exactamente 38,9%) son hombres y el 61% mujeres (exactamente 61,1%).

116. Se trataría, en definitiva, de elegir una magnitud que nos ayude a determinar "¿qué es ser alto?", "¿qué es ser rico?", "¿qué es ser gordo"... En la vida cotidiana normalmente definimos como altos a los que sobrepasan la "normalidad", esto es, el promedio de la población.
117. Téngase en cuenta que de las 20 respuestas obtenidas en la categoría "otras", 6 proceden de acudir a espectáculos (concretamente ir al teatro y a conciertos), y otras cuatro a "quedar" (con los amigos/as o con el novio/a).
118. Estas diferencias dependerán el tamaño muestral, puesto que con pequeños tamaños muestrales se produce un aumento del error típico de las estimaciones y una pérdida de significatividad de las diferencias porcentuales detectadas (Cea D'Ancona, 2004: 407). Reproducios una cita literal de López Pintor y Wert, por la precisión de la explicación: "...si la muestra tiene, como es usual, un margen de error del 3%, sólo tiene sentido empezar a pensar en diferencias por encima de los cuatro o cinco puntos de porcentaje, y en el caso de una distribución de frecuencias del conjunto de la muestra" (López Pintor y Wert, 2000: 537).

Tabla de contingencia (v01bis) Actividad, fuera de casa, que más te gusta hacer cuando dispones de tiempo libre * (v49) Género

Porcentajes de columna:

			(v49) Genero		
			Hombre	**Mujer**	**Total**
(v01bis) Actividad, fuera de casa...	Beber, ir de copas	Recuento	26	20	46
		% de (v49) Género	[119]**37,1%**	18,2%	25,6%
	Bailar	Recuento	0	12	12
		% de (v49) Género	,0%	**10,9%**	6,7%
	Hacer deporte	Recuento	12	8	20
		% de (v49) Género	**17,1%**	7,3%	11,1%
	Ir de excursión y al monte	Recuento	6	6	12
		% de (v49) Género	8,6%	5,5%	6,7%
	Viajar	Recuento	12	28	40
		% de (v49) Género	17,1%	25,5%	22,2%
	Ir al cine	Recuento	0	10	10
		% de (v49) Género	,0%	**9,1%**	5,6%
	Practicar alguna afición o hobby	Recuento	12	8	20
		% de (v49) Género	**17,1%**	7,3%	11,1%
	Otras	Recuento	2	18	20
		% de (v49) Género	2,9%	**16,4%**	11,1%
Total		Recuento	70	110	180
		% de (v49) Género	100,0%	100,0%	100,0%

Tabla 9.8. Porcentajes de la tabla 9.1 (parte 1).

119. Se han resaltado en negrilla los porcentajes superiores a la columna "total".

Porcentajes de fila:

			(v49) Genero		
			Hombre	**Mujer**	**Total**
(v01bis) Actividad, f uera de casa...	Beber, ir de copas	Recuento	26	20	46
		% de v01bis	[122]**56,5%**	*43,5%*	100,0%
	Bailar	Recuento	0	12	12
		% de v01bis	,0%	**100,0%**	100,0%
	Hacer deporte	Recuento	12	8	20
		% de v01bis	**60,0%**	*40,0%*	100,0%
	Ir de excursión y al monte	Recuento	6	6	12
		% de v01bis	50,0%	50,0%	100,0%
	Viajar	Recuento	12	28	40
		% de v01bis	*30,0%*	**70,0%**	100,0%
	Ir al cine	Recuento	0	10	10
		% de v01bis	,0%	**100,0%**	100,0%
	Practicar alguna afición o hobby	Recuento	12	8	20
		% de v01bis	**60,0%**	*40,0%*	100,0%
	Otras	Recuento	2	18	20
		% de v01bis	*10,0%*	**90,0%**	100,0%
Total		Recuento	70	110	180
		% de v01bis	38,9%	61,1%	100,0%

Tabla 9.8. Porcentajes de la tabla 9.1 (parte 2).

122. Se han resaltado en negrilla los porcentajes superiores a la fila "total" (última línea). Los subrayados indican porcentajes notablemente inferiores al total.

Porcentajes respecto del total:

			(v49) Genero		
			Hombre	**Mujer**	**Total**
(v01bis) Actividad, fuera de casa...	Beber, ir de copas	Recuento	26	20	46
		% del total	14,4%	11,1%	25,6%
	Bailar	Recuento	0	12	12
		% del total	,0%	6,7%	6,7%
	Hacer deporte	Recuento	12	8	20
		% del total	6,7%	4,4%	11,1%
	Ir de excursión y al monte	Recuento	6	6	12
		% del total	3,3%	3,3%	6,7%
	Viajar	Recuento	12	28	40
		% del total	6,7%	15,6%	22,2%
	Ir al cine	Recuento	0	10	10
		% del total	,0%	5,6%	5,6%
	Practicar alguna afición o hobby	Recuento	12	8	20
		% del total	6,7%	4,4%	11,1%
	Otras	Recuento	2	18	20
		% del total	1,1%	10,0%	11,1%
Total		Recuento	70	110	180
		% del total	38,9%	61,1%	100,0%

Tabla 9.8. Porcentajes de la tabla 9.1 (parte 3).

Con todos los porcentajes disponibles:

			(v49) Genero		
			Hombre	**Mujer**	**Total**
(v01bis) Actividad, fuera de casa...	Beber, ir de copas	Recuento	26	20	46
		% de v01bis	56,5%	43,5%	100,0%
		% de v49 Género	37,1%	18,2%	25,6%
		% del total	14,4%	11,1%	25,6%
	Bailar	Recuento	0	12	12
		% de v01bis	,0%	100,0%	100,0%
		% de v49 Género	,0%	10,9%	6,7%
		% del total	,0%	6,7%	6,7%
	Hacer deporte	Recuento	12	8	20
		% de v01bis	60,0%	40,0%	100,0%
		% de v49 Género	17,1%	7,3%	11,1%
		% del total	6,7%	4,4%	11,1%
	Ir de excursión y al monte	Recuento	6	62	
		% de v01bis	50,0%	50,0%	100,0%
		% de v49 Género	8,6%	5,5%	6,7%
		% del total	3,3%	3,3%	6,7%
	Viajar	Recuento	12	28	40
		% de v01bis	30,0%	70,0%	100,0%
		% de v49 Género	17,1%	25,5%	22,2%
		% del total	6,7%	15,6%	22,2%
	Ir al cine	Recuento	0	10	10
		% de v01bis	,0%	100,0%	100,0%
		% de v49 Género	,0%	9,1%	5,6%
		% del total	,0%	5,6%	5,6%
	Practicar alguna afición o hobby	Recuento	12	8	20
		% de v01bis	60,0%	40,0%	100,0%
		% de v49 Género	17,1%	7,3%	11,1%
		% del total	6,7%	4,4%	11,1%
	Otras	Recuento	2	18	20
		% de v01bis	10,0%	90,0%	100,0%
		% de v49 Género	2,9%	16,4%	11,1%
		% del total	1,1%	10,0%	11,1%
Total		Recuento	70	110	180
		% de v01bis	38,9%	61,1%	100,0%
		% de v49 Género	100,0%	100,0%	100,0%
		% del total	38,9%	61,1%	100,0%

Tabla 9.8. Porcentajes de la tabla 9.1 (parte 4).

El análisis *transversal*, que en este caso será comparando las filas con el total de fila (puesto que se trata de porcentajes de fila), desvela que los hombres emplean fundamentalmente su tiempo libre en beber (56,5% y 38,9%), hacer deporte (60% y 38,9%) y practicar alguna afición o hobby (60% y 38,9%); mientras que las actividades de ocio más elegidas por las mujeres son bailar (100% y 61,1%), viajar (70% y 61,1%), ir al cine (100% y 61,1%), y otras (75% y 61,1%). Ir de excursión y al monte es la única actividad que no presenta ninguna diferencia por sexo. Evidentemente cada investigador puede pedir la opción que le sea más cómoda, puesto que bastará con cambiar la posición de cada variable en las filas y en las columnas para que la tabla cambie de sentido. Si optamos por los porcentajes de fila, pero queremos obtener una tabla para comparar los hábitos de hombres y mujeres (v01bis) bastará con colocar esta variable en las filas, y la otra en las columnas en el cuadro de diálogo de la figura 9.1.

- Porcentajes totales calculados dividiendo el número de respuestas en cada celdilla entre el total de la tabla.
  Cálculo de la segunda columna: 20/180 = 0,111; 12/180 = 0,067; 8/180 = 0,044...
  El tercer ejemplo de la tabla 9.8 se interpreta utilizando conjuntamente la información de las filas y las columnas, es decir, un 14,4% de todos los entrevistados señalan que la actividad fuera de casa que más les gusta hacer cuando disponen de tiempo libre es beber, un 5,6% (de todos los entrevistados) apuestan por ir al cine, etc. En este caso el porcentaje "total de filas" corresponde a la frecuencia de filas (ver similitud con el primer ejemplo presentado en la tabla 9.8), mientras que el porcentaje total de columnas son las frecuencias de la variable columna.

El cálculo del porcentaje, como se ha expuesto, es tremendamente sencillo y no reviste complicación alguna. Algo más difícil es la decisión sobre que porcentajes son necesarios para realizar la interpretación de los datos. El criterio para esta decisión está condicionada por la formulación de la hipótesis utilizada, fruto de la *pregunta* que da lugar a la investigación o al marco teórico elegido. En este ejemplo, utilizado para conocer las actividades de ocio que caracterizan a los hombres y las mujeres, consideramos que resulta más adecuada la utilización de los porcentajes de columna en la medida que *elimina* el diferente número de hombres y mujeres seleccionadas. Aunque los resultados son muy similares, utilizar el porcentaje de filas puede generar algunas apreciaciones incorrectas[121], como tendremos ocasión de demostrar en el siguiente apartado.

121. Nos referimos, concretamente, a la actividad "viajar" que ha sido resaltada en el comentario de la tabla de los porcentajes de filas, pero no así en la de columnas.

Algunos investigadores (entre otros Cea D'Ancona, 2004: 407) recomiendan utilizar los porcentajes de columna[122] considerando su mayor facilidad de lectura; basados en el hecho que –en la cultura occidental– la lectura se realiza en sentido horizontal. Los porcentajes verticales, calculados respecto a la variable situada en la columna, se comparan entre ellos horizontalmente (de la misma forma en que se realiza la lectura). Otros preferimos los porcentajes de fila por la facilidad de comparar los porcentajes *de arriba a abajo* (en sentido vertical), así como por la *ausencia de limitación* del número de categorías presentes en la variable independiente. Ciertos investigadores colocan en columnas la variable con menos categorías, indicando en cada tabla el sentido de los porcentajes. A nuestro juicio esta opción puede desconcertar al lector al tener que cambiar –en cada tabla– el criterio de lectura. Se opte por uno u otro, siempre será necesario indicar claramente en la tabla el sentido de los porcentajes.

Lo que no recomendamos, bajo ningún concepto, es solicitar todos los porcentajes en una sola tabla, aún cuando tal solicitud se realice con el fin de decidir después cual utilizar. Solicitar los porcentajes de columna, fila y total dificulta tremendamente la interpretación de la tabla; como puede apreciarse en el análisis de la parte última de la tabla 9.8.

Antes de proceder con nuevos conocimientos recomendamos a los lectores interpretar el *interior* de las tablas elaboradas de los ejercicios realizados al final de la sección 9.3, esto es, relación entre el grado de felicidad (a54) y el estado civil (e12); grado de felicidad y sexo (e9); y relación entre clase social de pertenencia (e26) y nivel de equipamiento del hogar.

4.2. Interpretación de los valores de la tabla utilizando los residuos

El análisis de los *residuo*s supone una excelente opción para la interpretación de *tablas de contingencia*, y llama la atención la escasa utilización de esta estrategia en las investigación actual. Comenzaremos la exposición olvidándonos –por un momento– de los valores de las celdillas, considerando únicamente los *marginales* de la tabla. Nuestro interés es conocer cuales hubieran sido los valores de cada celdilla si únicamente conociéramos los valores marginales o, dicho de otro modo, qué valores podrían *esperarse* al considerar las filas y las columnas de la tabla. En esta situación podríamos calcular los valores de la celdilla multiplicando la frecuencia de la fila por la frecuencia de la columna, para dividirlo todo entre el total de la tabla; tal y como

122. Siempre que esta variable no presente un gran número de categorías de respuesta.

se expuso en la tabla 9.3. Recuérdese que estos valores se conocen con el nombre de *frecuencias esperadas* (o frecuencias teóricas), y –como ya indicamos– son las frecuencias que tendrían las celdillas si no existiera relación entre variables. Estas frecuencias esperadas se solicitan en el SPSS marcando la opción *Frecuencia esperada* en la parte superior izquierda del cuadro de diálogo mostrado en la figura 9.3.

Se ha afirmado que si no existiera ninguna relación entre las variables que forman la tabla las celdillas tendrían el valor de las frecuencias esperadas, es decir la diferencia entre frecuencias esperadas y observadas sería cero. Esto significa que cuanto mayor sea la diferencia entre las frecuencias esperadas y las obtenidas la relación entre las variables será mayor. La diferencia entre estas magnitudes recibe varios nombres, aunque en el ámbito de las tablas de contingencia se le conoce como *residuos*. El *residuo* es la diferencia entre la frecuencia esperada y la observada, tal y como se aprecia en los cálculos del cuadro 9.5 y en los resultados mostrados en la tercera fila de cada una de las celdillas de la tabla 9.9. Un *residuo* positivo significa que la frecuencia observada es mayor que la esperada, mientras que cuando es negativo indica lo contrario.

El problema al que nos enfrentamos ahora es cuantificar el valor 8,1 del *residual* de la celdilla de la primera fila y primera columna de la tabla 9.9 (marcado en negrilla) que, como se ha dicho, es la diferencia entre la frecuencia observada y esperada de la citada celdilla (26 – 17,9). Se trata de comparar los valores de cada *residuo* puesto que –como se aprecia en esta tabla– existe una gran variabilidad en el tamaño de cada uno. Supongamos que tuviéramos otra celdilla con un *residual* similar, pero que fuera obtenido de la resta 100008,1 - 100000. ¿Pueden interpretarse igual ambas magnitudes? En ambos casos la diferencia es la misma, en el primer caso la diferencia de 8,1 entre 26 y 18 es importante, pero en el segundo la diferencia de 8 entre 100008 y 100000 es ridícula. Es por ello por lo que es aconsejable utilizar los *residuo*s eliminando el efecto que puedan tener los marginales sobre su valor, residuos *estandarizados* según su frecuencia esperada. Son denominados como residuos *tipificados* o *estandarizados*. El análisis de la cuarta magnitud de cada una de las celdillas de la tabla 9.1 muestra un *alisamiento* de los valores de los residuos al ajustarlos, alisamiento que varía según el número de casos en los que se fundamenta cada residuo. Al eliminar el influjo del tamaño muestral ya es posible realizar comparaciones entre ellos: la mayor diferencia se produce entre las opciones bailar (residuo tipificado de hombres –2,2 y de mujeres 1,7) e ir al cine con valores *residual*es tipificados en los hombres de –2,0 y de mujeres de 1,6. Estas actividades son las que más diferencian y más definen las actividades de ocio (fuera del hogar) de hombres y mujeres.

Una mejor solución propone Haberman (1973) al dividir el residuo tipificado entre la raíz cuadrada de la varianza del residuo, calculada tal y como se muestra en la última parte del cuadro 9.5. Estos residuos se interpretan como cualquier

1.- **Frecuencias esperadas (o teóricas)** (FE)**:**

Columna 1:

70 * 46 / 180 = 17,9
70 * 12 / 180 = 4,7
70 * 20 / 180 = 7,8
70 * 12 / 180 = 4,7
70 * 40 / 180 = 15,6
70 * 10 / 180 = 3,9
70 * 20 / 180 = 7,8
70 * 20 / 180 = 7,8

2.- **Residuos** (FO-FT):

Columna 1:

26 - 17,9 = **8,1**
0 - 4,7 = -4,7
12 - 7,8 = 4,2
6 - 4,7 = 1,3
12 - 15,6 = -3,6
0 - 3,9 = -3,9
12 - 7,8 = 4,2
8 - 7,8 = -5,8

3.- **Residuos tipificados** (**estandarizados**): Res / √ FT

Columna 1:

8,1 / √ 17,9 = 1,9
-4,7 / √ 4,7 = -2,2
4,2 / √ 7,8 = 1,5
1,3 / √ 4,7 = 0,6
-3,6 / √ 15,6 = -0,9
-3,9 / √ 4,7 = -2,0
4,2 / √ 7,8 = 1,5
-5,8 / √ 7,8 = -2,1

4.- **Residuos tipificados ajustados o corregidos**:

Std Res / √ V_{ij} (Haberman, 1973: 215).

V_{ij} es una estimación de la varianza de e_{ij}, calculada con la expresión

V_{ij}: [1- (FO_i / n)] * [1- (FO_j / n)]

Donde:

FO_i: total de fila, frecuencia observada de fila
FO_j: total de columna, frec. observada de columna
n : tamaño muestral.

Cálculo:

(1 - [70/180]) * (1- [46/180]) = 0,4549
(1 - [70/180]) * (1- [12/180]) = 0,5704
(1 - [70/180]) * (1- [20/180]) = 0,5432
(1 - [70/180]) * (1- [12/180]) = 0,5703
(1 - [70/180]) * (1- [40/180]) = 0,4753
(1 - [70/180]) * (1- [10/180]) = 0,5772
(1 - [70/180]) * (1- [20/180]) = 0,5432
(1 - [70/180]) * (1- [20/180]) = 0,5432

Cálculo de residuos ajustados:

1,9 / √ 0,4549 = 2,81
-2,2 / √ 0,5704 = -2,91
1,5 / √ 0,5432 = 2,09
0,6 / √ 0,5703 = 0,79
-0,9 / √ 0,4753 = 1,30
-2,0 / √ 0,5772 = -2,63
1,5 / √ 0,5432 = 2,09
2,1 / √ 0,5432 = 2,85

Nota: a fin de simplificar los cálculos se ha calculado únicamente en la columna de los hombres. El proceso es elmismo para el resto de las celdillas.

Cuadro 9.5. Cálculo de los componentes de una tabla de contingencia (ejemplo con la tabla 9.1).

Tabla de contingencia (v01bis) Actividad, fuera de casa, que más te gusta hacer cuando dispones de tiempo libre * (v49) Género

			(v49) Genero		
			Hombre	**Mujer**	**Total**
(v01bis) Actividad, fuera de casa, que más te gusta hacer cuando dispones de tiempo libre	Beber, ir de copas	Recuento	26	20	46
		Frec. esperada	17,9	28,1	46,0
		Residuo	**8,1**	-8,1	
		Residuo tipificado	1,9	-1,5	
		Residuo corregido	2,8	-2,8	
	Bailar	Recuento	0	12	12
		Frec. esperada	4,7	7,3	12,0
		Residuo	-4,7	4,7	
		Residuo tipificado	-2,2	1,7	
		Residuo corregido	-2,9	2,9	
	Hacer deporte	Recuento	12	8	20
		Frec. esperada	7,8	12,2	20,0
		Residuo	4,2	-4,2	
		Residuo tipificado	1,5	-1,2	
		Residuo corregido	2,1	-2,1	
	Ir de excursión y al monte	Recuento	6	6	12
		Frec. esperada	4,7	7,3	12,0
		Residuo	1,3	-1,3	
		Residuo tipificado	,6	-,5	
		Residuo corregido	,8	-,8	
	Viajar	Recuento	12	28	40
		Frec. esperada	15,6	24,4	40,0
		Residuo	-3,6	3,6	
		Residuo tipificado	-,9	,7	
		Residuo corregido	-1,3	1,3	
	Ir al cine	Recuento	0	10	10
		Frec. esperada	3,9	6,1	10,0
		Residuo	-3,9	3,9	
		Residuo tipificado	-2,0	1,6	
		Residuo corregido	-2,6	2,6	
	Practicar alguna afición o hobby	Recuento	12	8	20
		Frec. esperada	7,8	12,2	20,0
		Residuo	4,2	-4,2	
		Residuo tipificado	1,5	-1,2	
		Residuo corregido	2,1	-2,1	
	Otras	Recuento	2	18	20
		Frec. esperada	7,8	12,2	20,0
		Residuo	-5,8	5,8	
		Residuo tipificado	-2,1	1,7	
		Residuo corregido	-2,8	2,8	
Total	Recuento	70	110	180	
	Frec. esperada	70,0	110,0	180,0	

Tabla 9.9. Ejemplo de tabla con todas las opciones disponibles en la opción *Casillas*.

valor de una variable tipificada (estandarizada) con una distribución normal: un valor superior a +1,96, o inferior a –1,96, indica que hay relación entre ambas categorías a un *nivel de confianza* del 95%, y el +/-2,58[123] indica que existe relación a un nivel de confianza del 99% (Haberman 1973: 216-218). De este modo cuanto mayor es el valor del residuo mayor es la diferencia, indicando el signo la dirección de la relación.

De la tabla 9.9 se extrae la 9.10 donde se muestra en cada celdilla únicamente el número de casos y los residuos tipificados corregidos. Basta con un breve vistazo a esta tabla para detectar las mayores relaciones, aquellas celdillas donde se encuentran los residuos con mayor valor (siempre que sean superiores a 1,96 y menores de –1,96, que son los umbrales de *significación* al 95%). En la tabla 9.10 los residuos significativos se han destacado en negrilla para facilitar la interpretación.

Los dos primeras filas presentan las magnitudes más altas, lo que indica una gran relación entre variables, relación que alcanza su punto álgido entre los entrevistados que eligen bailar como la actividad fuera de casa que más hacen en su tiempo libre; con un valor –2,9 para los hombres y de 2,9 para las mujeres[124]. El valor de estos *residual*es está indicando que existe una gran relación (positiva) entre las mujeres y bailar, y negativa entre los hombres y bailar. Dicho de otro modo, las mujeres destacan por emplear su tiempo de ocio bailando, mientras que los hombres destacan por las pocas elecciones en esta actividad. El alto valor positivo del residuo correspondiente a la celdilla hombres y beber está indicando que los hombres destacan por emplear el ocio en esta actividad, caso contrario al de las mujeres.

Similar interpretación cabe hacer de la práctica del deporte y de practicar alguna actividad o hobby, actividades elegidas fundamentalmente por el colectivo masculino. Las mujeres, además de bailar, destacan también por acudir al cine. En definitiva, una interpretación idéntica a la realizada en el comentario a la tabla 9.8 (tabla con porcentajes de columna y fila), si bien –desde mi punto de vista– con una mayor sencillez puesto que no hay que decidir el *tipo* de porcentajes a utilizar, como compararlos, etc. Es importante destacar la nula referencia a viajar, puesto que presenta unos residuos corregidos no significativos. Recuérdese, como se señaló en la nota a pie número 20, que esta actividad únicamente apareció en el comentario de los porcentajes de filas, y no en los porcentajes verticales (de columnas).

123. Son los *valores críticos* estadístico que indican las zonas de significación de la curva normal, con un nivel de confianza del 95 y del 99% respectivamente (Everitt y Wykes, 2001: 211).

124. Téngase en cuenta que en el caso de una tabla con dos columnas, como ocurre en este ejemplo, los residuos se *contraponen* con el fin de sumar 0. El análisis de los residuos corregidos es más interesante con tablas mayores, en aquellas donde el análisis de los porcentajes resulta más complicado.

Tabla de contingencia (v01bis) Actividad, fuera de casa, que más te gusta hacer cuando dispones de tiempo libre * (v49) Género

			(v49) Genero		
			Hombre	**Mujer**	**Total**
(v01bis) Actividad, fuera de casa, que más te gusta hacer cuando dispones de tiempo libre	Beber, ir de copas	Recuento	26	20	46
		Residuo corregido	**2,8**	**-2,8**	
	Bailar	Recuento	0	12	12
		Residuo corregido	**-2,9**	**2,9**	
	Hacer deporte	Recuento	12	8	20
		Residuo corregido	**2,1**	**-2,1**	
	Ir de excursión y al monte	Recuento	6	6	12
		Residuo corregido	,8	-,8	
	Viajar	Recuento	12	28	40
		Residuo corregido	-1,3	1,3	
	Ir al cine	Recuento	0	10	10
		Residuo corregido	**-2,6**	**2,6**	
	Practicar alguna afición o hobby	Recuento	12	8	20
		Residuo corregido	**2,1**	**-2,1**	
	Otras	Recuento	2	18	20
		Residuo corregido	**-2,8**	**2,8**	
Total	Recuento	70	110	180	

Tabla 9.10. Ejemplo de tabla con residuos tipificados corregidos.

Antes de terminar con la explicación de los *residual*es señalar que el procedimiento propuesto por Haberman ha sido utilizado en la investigación con encuesta en nuestro país en los trabajos de Felman y otros (1988), González (1994), Ayerdi (1995) y Díaz de Rada (2004).

Del cuadro de diálogo principal mostrado en la figura 9.1 tan sólo resta explicar el botón *Exactas...*, que supera ampliamente los objetivos aquí planteados, y *Formato...*, que está referido al orden de presentación de la tabla. El cuadro de diálogo que surge tras pulsar este botones se muestra en la figura 9.4 y permite presentar las filas en orden ascendente o descendente; siempre considerando la codificación de la variable colocada en filas (no el número de casos contados). Por defecto aparece la presentación en orden ascendente (figura 9.4); esto es del valor inferior al superior, que es como se han mostrado las tablas presentadas hasta el momento: recuérdese que la opción *beber ir de copas* estaba codificada con el valor 1, *bailar* con el 2, etc. Por último, y vol-

viendo a la figura 9.1, obsérvese que en la parte inferior izquierda aparecen dos opciones que permiten *suprimir tablas*, y *mostrar los gráficos de barras agrupadas*. Disponer de gráficos supone una notable ayuda en la interpretación de la tabla de contingencia; si bien consideramos que son más interesantes los gráficos mostrados desde el menú gráficos. Por ello recomendamos no ejecutar los gráficos desde aquí puesto que se realizan en base a las frecuencias observadas (números absolutos), y no permite la utilización de porcentajes. Estos motivos nos llevan a recomendar el menú Gráficos (mostrado en la figura 4.9); que además presenta un mayor número de dispositivos gráficos.

Fijar los conocimientos aprendidos es el objetivo de toda actividad docente. Con el fin de consolidar lo aprendido recomendamos interpretar el interior de la tabla de los ejercicios planteados al final de la sección anterior. Se trataba, concretamente, de:

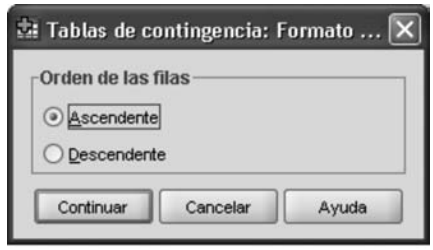


Figura 9.4. Botón *Formato* dentro de cuadro de diálogo de la Tablas de contingencia. Mostrar interior de la tabla.

- Relación entre grado de felicidad (a54) y sexo (e9); entre grado de felicidad (a54) y estado civil (e12).
- Relación entre clase social de pertenencia (e26) y "nivel de equipamiento del hogar" (variable creada en el apartado 8.7 en base a la pregunta 20).

5. Utilización de test estadísticos para conocer la relación entre variables ordinales

Todos los ejemplos presentados hasta ahora se han referido a variables nominales, buscando satisfacer los objetivos de la investigación planteada; que recordemos se fundamentaba en conocer las actividades de ocio (fuera del hogar) que caracterizan a los hombres y a las mujeres. Resuelta esa cuestión, en este momento plantea-

mos si existe relación entre el número de libros leídos en los últimos cinco meses no relacionados con los estudios (v08) y los libros leídos relacionados con los estudios (v06). Es decir, se trata de determinar si los entrevistados que presentan un mayor índice de lectura de libros (en general) son también los que más leen libros relacionados con tus estudios, o si más bien se trata del efecto contrario: si los que leen todo tipo de libros presentan un menor nivel de lectura de libros relacionados con los estudios. De modo que la *pregunta de investigación* que guiará la exposición en este apartado propone lo siguiente: ¿podríamos decir que los entrevistados que más leen todo tipo de libros[125] son también los que leen más libros relacionados con sus estudios?; o –más bien– al leer otros libros reducen la lectura de textos relacionados con sus estudios. (Dicho de otro modo, ¿existe relación entre el número de libros *relacionados* con los estudios y la lectura de libros *generalistas*?).

Las frecuencias de ambas variables, mostradas en la tabla 9.11, desvelan que un 20% de los entrevistados no leen ningún libro, considerando tanto los relacionados con los estudios como los no relacionados. Las frecuencias de la izquierda dan cuenta de los libros leídos relacionados con los estudios, y muestra que uno de cada cuatro entrevistados lee un libro, un 15% dos libros, y el 14% tres libros. Es interesante utilizar la columna porcentaje acumulado para conocer el número de entrevistados que leen tres y menos libros, estrategia que requiere tener en cuenta que la categoría *ninguno* aparece en la primera fila de la tabla; de modo que será necesario restar al porcentaje acumulado (73,8%) el porcentaje de personas que no han leído ningún libro (20,4%). De este modo obtenemos que el 53,4% de los entrevistados (73,8 – 20,4) han leído menos de cuatro libros.

Es posible emplear una estrategia que evite realizar esta resta, colocando la categoría *ninguno* en la parte superior de la distribución, esto es, codificándola con un valor mayor al valor más alto de la distribución. En el caso de v06, por ejemplo, podríamos codificar esta categoría con el valor 50[126], de modo que este categoría quedará situada por encima del 15, facilitando la lectura del porcentaje acumulado. Sin embargo, no recomendamos esta práctica puesto –como indicamos en la sección 3.4– "la medición ordinal debe respetar las relaciones observadas en la asignación del sistema de medición, ordenando los números según su orden serial". Codificar la categoría el ninguno con el valor 50, además de *romper* el orden de la distribución (al pasar del 15 al 50), altera su orden serial que –recordemos– implica que los valores más altos de la distribución son –en este caso– los que más libros leen.

125. A partir de ahora, y con el fin de simplificar la exposición, nos referiremos a éstos como *libros de todo tipo* o libros *generalistas*.

126. Indicando, en el procedimiento *Recodificar en distintas variables*", que el valor 0 es igual a 50. Posteriormente habría que *etiquetar* el valor 50 con la opción *ninguno*.

V06 Libros relacionados con tus estudios leídos en los últimos cinco meses				v08 libros leídos en los últimos cinco meses no relacionados con tus estudios (libros generalistas)			
	Frec.	Porcent	Porcent acum		Frec.	Porcent	Porcent acum
Ninguno	39	20,4	20,4	Ninguno	38	19,9	19,9
1	48	25,1	45,5	1	34	17,8	37,7
2	28	14,7	60,2	2	32	16,8	54,5
3	26	13,6	73,8	3	34	17,8	72,3
4	12	6,3	80,1	4	10	5,2	77,5
5	8	4,2	84,3	5	6	3,1	80,6
6	12	6,3	90,6	6	4	2,1	82,7
7	4	2,1	92,7	10	10	5,2	88,0
8	4	2,1	94,8	11	2	1,0	89,0
10	6	3,1	97,9	12	2	1,0	90,1
11	2	1,0	99,0	14	2	1,0	91,1
15	2	1,0	100,0	15	6	3,1	94,2
			20	2	1,0	95,3	
			20	2	1,0	96,3	
				No responde	7	25,1	100,0
Total	191	100,0		Total	191	100,0	

Tabla 9.11. Frecuencias de v06 y v08.

El análisis de v08, donde se recoge el número de libros leídos no relacionados con los estudios (libros generalistas), presenta 7 personas que no responden, de modo que debe definirse ese valor como *perdido* con el fin de poder interpretar adecuadamente el porcentaje de respuestas obtenidas. En la tabla 9.12 se presenta la tabla de frecuencias sin considerar las no respuestas, que muestran unas cifras ligeramente superiores a las mostradas en la variable v06: un 18,5% de los entrevistados ha leído un libro generalista, el 17% dos, y el 18% tres. El análisis del porcentaje acumulado (restando el porcentaje de la categoría *ninguno*) desvela que un 54% de los entrevistados han leído menos de cuatro libros generalistas. Esta similitud en los porcentajes nos lleva a sospechar que los entrevistados que presentan un mayor índice de lectura de libros (en general) son

(v08) Libros leídos en los últimos cinco meses no relacionados con los estudios

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Ninguno	38	19,9	20,7	20,7
	1	34	17,8	18,5	39,1
	2	32	16,8	17,4	56,5
	3	34	17,8	18,5	75,0
	4	10	5,2	5,4	80,4
	5	6	3,1	3,3	83,7
	6	4	2,1	2,2	85,9
	10	10	5,2	5,4	91,3
	11	2	1,0	1,1	92,4
	12	2	1,0	1,1	93,5
	14	2	1,0	1,1	94,6
	15	6	3,1	3,3	97,8
	20	2	1,0	1,1	98,9
	24	2	1,0	1,1	100,0
	Total	184	96,3	100,0	
Perdidos	No responde	7	3,7		
Total	191	100,0			

Tabla 9.12. Frecuencias de v08 considerando las no respuestas (valor 99) como valor perdido.

también los que más leen libros relacionados con tus estudios. No obstante, será necesario utilizar el cruce de tablas para dar respuesta a esta hipótesis.

Buscando complicar aún más nuestra exposición –y con el fin de repasar procedimientos vistos anteriormente– supondremos que el demandante de la investigación ha indicado que desea conocer la relación de v06 con v08 únicamente entre los entrevistados que menos leen, es decir, en aquellos que leen 0, 1, 2 y 3 libros.

Para realizar los análisis en este colectivo utilizaremos la selección de casos mediante criterios condicionales que fue explicada en la sección 8.8. Tras pulsar *Datos⇒Seleccionar Casos* se marca la opción *Si se satisface la condición* (figura 8.20) y aparece el cuadro de diálogo que se muestra en la figura 9.5. Seleccionada la variable v06, se desplaza a la ventana central para añadir los símbolos que indican *menor o igual*

a tres. Posteriormente se incluye el operando y, se desplaza la variable v08 a la ventana central, y se indica –en este caso– que seleccione los valores menores al cuatro[127]. Pulsando el botón *Continuar* y *Aceptar* se lleva a cabo la selección de los entrevistados que leen menos de cuatro libros.

Figura 9.5. Cuadro de diálogo *Seleccionar casos*, condición lógica con dos términos.

Será necesario solicitar las frecuencias de ambas variables para conocer cómo la selección efectuada afecta a ambas distribuciones. En la tabla 9.12 puede apreciarse una reducción en el tamaño muestral de 191 a 112 casos, así como los escasos cambios (en ambas variables) en la categoría *ninguno*. Comparando la distribución obtenida con la mostrada en la tabla 9.11 podemos apreciar –en v06– una disminución en todas las categorías: los entrevistados que no leen ningún libro descienden de 39 a 36, aquellos que leen un libro se reducen de 48 a 42, los que leen dos libros de 28 a 18, y los lectores de tres libros de 26 a 16. Menores cambios se producen en v08 puesto que el número de entrevistados que no lee ningún libro se mantiene estable, y se reduce ligeramente aquellos que leen un libro (de 34 a 32). En el resto de categorías de v08 se produce una reducción similar a la experimentada en v06: el número de los entrevistados que leen dos libros disminuye de 32 a 20, y los que leen tres libros de 34 a 22.

Hemos prestado atención al cambio entre ambas distribuciones para que el lector reflexione sobre las personas que no han sido seleccionadas, que son otros sino aquellos que leen muchos libros *generales* (altos valores en v08) y –a la vez– leen muchos libros relacionados con los estudios (altos valores en v06). Se trata de un colectivo de 77 entrevistados (191 – 112), un 40,3% (77 / 191 * 100) de la muestra original.

El demandante de la investigación indica que su interés de centra en conocer los que no leen ningún libro, los que leen uno, y –de forma agregada– los que leen dos

127. Seleccionar *menor o igual a tres* y *menor que cuatro* proporciona los mismos resultados puesto que el valor cuatro no está incluido en ninguna de las dos instrucciones. Se ha seleccionado *menor que cuatro* en v08 para mostrar el mayor número de recursos susceptibles de ser utilizados.

V06 Libros relacionados con tus estudios leídos en los últimos cinco meses				v08 libros leídos en los últimos cinco meses no relacionados con tus estudios			
	Frec.	Porcent.	Porcent acum		Frec.	Porcent valid	Porcent acum
Ninguno	36	32,1	32,1	Ninguno	38	33,9	33,9
1	42	37,5	69,6	1	32	28,6	62,5
2	18	16,1	85,7	2	20	17,9	80,4
3	16	14,3	100,0	3	22	16,9	100,0
Total	112	100,0		Total	112	100,096,3	
				No responde	7	25,1	100,0
Total	191	100,0		Total	191	100,0	

Tabla 9.13. Frecuencias de v06 y v08, seleccionados los entrevistados que leen tres y menos libros.

y tres libros. Aunque la experiencia investigadora recomienda siempre presentar los resultados de la forma más desagregada posible, siguiendo las indicaciones del demandante de la investigación se han agrupado las personas que leen dos y tres libros en (tanto en v06 como en v08). No reproducimos la tabla por motivos de espacio, y porque el lector puede construir fácilmente la tabla uniendo las categorías 2 y 3 de la tabla 9.13.

Con el fin de resolver el objetivo planteado[128] será necesario realizar una *tabla de contingencia* entre ambas variables siguiendo las instrucciones presentadas en el apartado 9.2, pero esta vez colocando v08 en columnas y v06 en filas. Solicitaremos también los porcentajes de columnas y los residuos corregidos. Posteriormente será necesario seleccionar las medidas de asociación precisas para conocer la relación entre ambas variables.

Antes de elegir las medidas de asociación para conocer la relación entre las variables será preciso preguntarnos por la escala de medida de éstas, que como se ha expuesto anteriormente (capítulo III) se trata de variables ordinales. Hay varias medidas que permiten analizar la relación entre variables ordinales, si bien aquí únicamente expondremos los más utilizados: *Gamma*, *d de Somer*, *Tau-b de Kendall* y *Tau-c de Kendall*. Basta con observar de nuevo el cuadro de diálogo *Estadísticos...*, presentado

128. Descubrir si existe relación entre el número de libros relacionados con los estudios, y la lectura de libros generalistas.

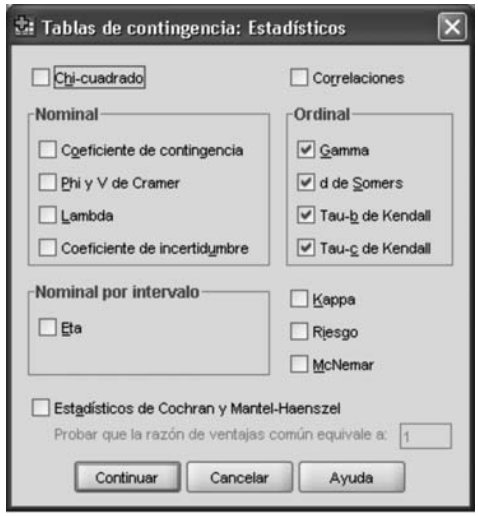


Figura 9.6. *Estadísticos* de tablas de contingencia para variables ordinales.

en la figura 9.6, para constatar que éstos se utilizan con variables ordinales. La tabla 9.14 muestra los resultados obtenidos para el cruce de v06 y v08.

El objetivo de la utilización de estadísticos para variables ordinales es analizar la relación entre la clasificación de cada individuo en cada una de las variables utilizadas; de modo que existirá relación cuando la distribución de los casos de la primera variable permita predecir la ordenación de los casos en la segunda variable (García Ferrando, 1985: 244). Nuestro objetivo se centra más en la interpretación de los coeficientes obtenidos que el cálculo de los mismos, de modo que –como en anteriores ocasiones–centraremos la exposición en la interpretación de los coeficientes, recomendando al lector interesado en los cálculos la consulta del anexo 1 donde se expone el cálculo de pares. La formulación de cada coeficiente se muestra en el cuadro 9.6.

El primero de los estadísticos solicitados, el coeficiente *Gamma* de Goodman y Kruskal es una medida simétrica que va desde –1 a 1 y se calcula restando al número de pares concordantes (N_c) los pares discordantes (N_d), y dividiendo este resultados entre la suma de ambos. Del cálculo de pares[129] se desprende que cuando todos los pares son concordantes existirá una asociación positiva entre las variables, puesto que los entrevistados que leen más libros relacionados con tus estudios serán tam-

129. Ver el anexo 1 donde se realiza una introducción al cálculo de pares.

A) Gamma de Goodman y Kruskal:

$$\text{Gamma} = \frac{N_c - N_d}{N_c + N_d} = \frac{1.828 - 1.064}{1.828 + 1.064} = 0{,}264$$

B) Tau-b de Kendall:

$$T_b = \frac{N_s - N_d}{\sqrt{(N_s + N_d + T_y)\,(N_s + N_d + T_x)}} =$$

$$T_b = \frac{1.828 - 1.064}{\sqrt{(1.828 + 1.064 + 1.264)\,(1.828 + 1.064 + 1.272)}} = 0{,}1836$$

donde T_x = son los pares empatados en la variable x, y
T_y = son los pares empatados en la variable Y.

C) Tau-c de Kendall:

$$T_c = \frac{2 * m * (N_s - N_d)}{N^2 * (m-1)} = \frac{2 * 3 * (1.828 - 1.064)}{112^2 * (3-1)} = 0{,}1827$$

donde m = mínimo número de filas o columnas en la tabla de contingencia
N = tamaño de la tabla

D) D de Somer:

$$d_{yx} = \frac{N_s - N_d}{N_s + N_d + T_y} = \frac{1.828 - 1.064}{1.828 + 1.064 + 1.264} = 0{,}1838$$

$$d_{xy} = \frac{N_s - N_d}{N_s + N_d + T_x} = \frac{1.828 - 1.064}{1.828 + 1.064 + 1.272} = 0{,}1834$$

Nota: el cálculo de pares se ha presentado en el anexo 1.

Cuadro 9.6. Medidas para conocer la relación entre variables ordinales (ejemplo con la tabla 9.15).

bién los que leen más libros no relacionados, mientras que un mayor número de pares discordantes provocan ausencia de relación, es decir que una persona que lee muchos libros (de todo tipo) leerá pocos libros relacionados con sus estudios. De este modo *Gamma* es el exceso de pares concordantes en relación al número de pares concordantes y discordantes, o dicho de otra forma es el número de predicciones correctas (número de entrevistados con el mismo orden en las dos variables) menos las incorrectas, dividido entre el total de predicciones. Tras constatar que se trata de un valor significativo, se procede con su interpretación. Un valor *Gamma* de 0,264 indica un 26% más de pares concordantes que discordantes, una mejora en la predic-

Medidas simétricas

		Valor	Error típ. sint.(a)	T aproximada (b)	Sig. aproximada
Ordinal por ordinal	Tau-b de Kendall	,184	,087	2,118	,034
	Tau-c de Kendall	,183	,086	2,118	,034
	Gamma	,264	,122	2,118	,034
N de casos válidos		112			

a Asumiendo la hipótesis alternativa.
b Empleando el error típico asintótico basado en la hipótesis nula.

Medidas direccionales

			Valor	Error típ. sint.(a)	T aproximada (b)	Sig. aproximada
Ordinal por ordinal	d de Somer	Simétrica	,184	,087	2,118	,034
		(v06) Libros relacionados con tus estudios leídos en los últimos cinco meses dependiente	,184	,087	2,118	,034
		(v08) Libros leídos en los últimos cinco meses no relacionados con los estudios dependiente	,183	,087	2,118	,034

a Asumiendo la hipótesis alternativa.
b Empleando el error típico asintótico basado en la hipótesis nula.

Tabla 9.14. *Estadísticos* para variables ordinales: D de Somer, Gamma, Tabu-b de Kendall y tau-c de kendall.

ción, una reducción del 26% del error al predecir los casos de una variable conociendo la ordenación de los casos en otra variable.

La mayoría de los expertos señalan que pese a que *Gamma* es medida muy versátil que es utilizada muy frecuentemente tiene el problema que suele sobreestimar la relación existente entre las variables analizadas, fundamentalmente si el número de pares empatados en una variable es muy elevado (Dometrius 1992: 309). Debido a estas

críticas algunos investigadores utilizan la *Tau-b* de Kendall que oscila entre –1 a 1 y su magnitud indica el grado de asociación entre dos variables: en que medida el cambio en una variable provocará cambios en la otra.

Esta medida tiene el inconveniente que únicamente puede utilizarse con tablas cuadradas (2 x 2, 3 x 3, 4 x 4, etc.) puesto que si las tablas no son cuadradas no puede llegar a uno, ya que cuando hay un número diferente de filas que de columnas existen más pares empatados en una variable que en la otra. Es por ello por lo que en tablas rectangulares se sustituye por *Tau-c*, que oscila entre –1 a 1 indicando su magnitud el grado de asociación entre las variables; esto es, cómo la variación de una variable produce variaciones en la segunda.

El siguiente estadístico, la *D de Somer* es una medida asimétrica que se interpreta de modo similar a *Gamma*, si bien presenta la ventaja que no sobreestima la relación entre variables al eliminar la influencia de los pares empatados en la variable dependiente. García Ferrando (1985: 253) explica la relación entre *D* y las *Taus* y afirma que la *tau-b* es un promedio de los dos coeficientes *D* de Somer que pueden calcularse en una misma tabla. Así la *D* es la reducción proporcional en el error cometido al predecir el ordenamiento de los casos en una variable mediante el conocimiento de la ordenación de los casos en otra variable, es decir que tiene una interpretación similar a la Gamma.

¿Que estadístico elegir?

Tras esta exposición, y considerando que todas las medidas explicadas son adecuadas para conocer la relación entre variables ordinales nos preguntamos qué estadístico es el mejor para conocer la relación entre dos variables. Ruiz Maya (1990: 287) considera que *Gamma* de Goodman y Kruskall es la más utilizada debido fundamentalmente a la facilidad de su interpretación, aunque cuando el número de empates es muy alto aconseja utilizar otras medidas. Dometrius (1992: 313) llega a una conclusión similar cuando expone que *Gamma* es más simple de calcular y de interpretar, aunque también afirma que la elección de la medida es un criterio personal y aconseja que cada investigador utilice la medida que mejor comprenda, opinión compartida también por Manzano (1995: 255). Dometrius, por otra parte, aconseja utilizar la *Tau* de Kendall porque es una medida más conservadora al eliminar el efecto de los empates, aunque señala que cuando el número de empates es pequeño el valor de la *Tau* será similar a *Gamm*a (992: 314).

Desde nuestro punto de vista creemos que la mejor medida es la *D de Somer* porque recoge la facilidad de interpretación de *Gamm*a, al tiempo que elimina el principal defecto de ésta (la sobreestimación de la relación) al excluir el efecto de los pares empatados. Además de ser la única medida asimétrica. No obstante nuestro consejo es que cada uno utilice la medida que mejor comprenda.

Tabla de contingencia (v08) Libros leídos en los últimos cinco meses no relacionados con los estudios * (v06) Libros relacionados con tus estudios leídos en los últimos cinco meses

			(v08) Libros leídos NO relacionados con los estudios (libros generalistas)			
			Ninguno	**Uno**	**Dos y tres**	**Total**
(v06) Libros relacionados con...	Ninguno	Recuento	22	2	12	36
		% de (v08) Libros NO relacionados con...	57,9%	6,3%	28,6%	32,1%
		Residuos corregidos	4,2	-3,7	-,6	
	Uno	Recuento	10	14	18	42
		% de (v08) Libros NO relacionados con...	26,3%	43,8%	42,9%	37,5%
		Residuos corregidos	-1,8	,9	,9	
	Dos y tres	Recuento	6	16	12	34
		% de (v08) Libros NO relacionados con...	15,8%	50,0%	28,6%	30,4%
		Residuos corregidos	-2,4	2,9	-,3	
Total		Recuento	38	32	42	112
		% de (v08) Libros NO relacionados con...	100,0%	100,0%	100,0%	100,0%

Tabla 9.15. Tabla de contingencia v08 y v06.

Otra cuestión a plantear es como interpretar la magnitud de cada estadístico. Dometrius (1992: 314) considera que magnitudes superiores al 0,3 ya indican niveles de asociación importantes. No obstante, desde su punto de vista una relación puede ser considerada fuerte o débil no tanto por ella misma sino en relación con el marco teórico previo y otras investigaciones similares. Este autor pone el ejemplo de un grado de asociación entre el *Nivel de Estudios* y el *Nivel de Ingresos* de 0,25. Éste será un importante resultado para analizar que elementos están detrás de este bajo nivel de asociación en unas variables que en nuestro (supuesto) marco teórico aparecían muy relacionadas.

En cualquier caso, se opte por una u otra medida, al observar el ejemplo utilizado se detecta que existe una relación *directa* entre el número de libros leídos en los últimos cinco meses y el número de libros leídos relacionados con tus estudios. Una relación directa implica que a medida que aumenta una variable se incrementa también los valores de la otra; es decir, que las personas que más leen todo tipo de libros son también las que leen más libros relacionados con sus estudios.

Resumen del procesamiento de los casos

	Casos					
	Válidos		Pérdidos		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
(v01bis) Actividad, fuera de casa, que más te gusta hacer cuando dispones de tiempo libre * (v49) Género	180	94,2%	11	5,8%	191	100,0%

Esta relación se observa con precisión en el interior de la tabla 9.15, bien utilizando los porcentajes de columna o los residuos corregidos. En este caso se han utilizado los porcentajes de columna, que indican que el 57,9% de los entrevistados que no leen libros generalistas tampoco leen otro tipo de libros. Sin embargo también es verdad que la mitad de los que han leído un libro generalista han leído dos o tres libros relacionados con los estudios, y un 44% un libro. De los que han leído *dos y tres* libros no relacionados con sus estudios (libros generalistas), el 43% ha leído un libro relacionado, y el 29% dos o tres libros.

Finalizaremos esta sección indicando que cuando se considera la relación entre una variable nominal y otra ordinal, como sucede en los ejercicios planteados dos párrafos más arriba, deben utilizarse los estadísticos y medidas de asociación propias de las variables nominales. Debe tenerse en cuenta que el programa calcula todo, absolutamente todo, y que el investigador es quién debe elegir qué interpretar en función de su hipótesis de trabajo y del tipo de variables que utiliza. En la tabla 9.16 se han solicitado todas las medidas de asociación vistas a lo largo de la exposición, medidas para variables nominales, para variables ordinales, e incluso para variables de intervalo[130]. El programa lo calcula absolutamente todo, y es el investigador el que debe discernir lo que debe interpretar.

Con el fin de fijar los conocimientos aprendidos en esta sección, antes de considerar nuevos contenidos, proponemos unos ejercicios utilizando el archivo de datos sobre *Vida Cotidiana*. ¿Hasta que punto el número de amigos de verdad (pregunta 34b, variable b68) está relacionado con la edad (e10) de los entrevistados (edad en cuatro

130. Obsérvese que en la tabla 9.16 aparecen dos medidas no tratadas, la *Correlación de Pearson* (R de Pearson) y *Correlación de Spearman*, no explicadas en este capítulo al alejarse de nuestros propósitos.

Tabla de contingencia (v01bis) Actividad, fuera de casa, que más te gusta hacer cuando dispones de tiempo libre * (v49) Género

			(v49) Genero		
			Hombre	**Mujer**	**Total**
(v01bis) Actividad, fuera de casa, que más te gusta hacer cuando dispones de tiempo libre	Beber, ir de copas	Recuento	26	20	46
		% de (v49) Género	37,1%	18,2%	25,6%
		Residuos corregidos	2,8	-2,8	
	Bailar	Recuento	0	12	12
		% de (v49) Género	,0%	10,9%	6,7%
		Residuos corregidos	-2,9	2,9	
	Hacer deporte	Recuento	12	8	20
		% de (v49) Género	17,1%	7,3%	11,1%
		Residuos corregidos	2,1	-2,1	
	Viajar	Recuento	12	28	40
		% de (v49) Género	17,1%	25,5%	22,2%
		Residuos corregidos	-1,3	1,3	
	Ir al cine	Recuento	0	10	10
		% de (v49) Género	,0%	9,1%	5,6%
		Residuos corregidos	-2,6	2,6	
	Practicar alguna afición o hobby	Recuento	12	8	20
		% de (v49) Género	17,1%	7,3%	11,1%
		Residuos corregidos	2,1	-2,1	
	Otras	Recuento	8	24	32
		% de (v49) Género	11,4%	21,8%	17,8%
		Residuos corregidos	-1,8	1,8	
Total	Recuento	70	110	180	
		% de (v49) Género	100,0%	100,0%	100,0%

Pruebas de chi-cuadrado

	Valor	**gl**	**Sig. asintótica (bilateral)**
Chi-cuadrado de Pearson	31,447(a)	6	,000
Razón de verosimilitud	38,885	6	,000
Asociación lineal por lineal	1,249	1	,264
N de casos válidos	180		

a 2 casillas (14,3%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 3,89.

Medidas direccionales

			Valor	Error típ. sint.(a)	T aproximada (b)	Sig. aproximada
Ordinal por ordinal	d de Somer	Simétrica	,141	,063	2,236	,025
		(v01bis) Actividad, fuera de casa, que más te gusta hacer... dependiente	,193	,086	2,236	,025
		(v49) Género dependiente	,112	,050	2,236	,025

a Asumiendo la hipótesis alternativa.
b Empleando el error típico asintótico basado en la hipótesis nula.

Medidas simétricas

		Valor	Error típ. sint.(a)	T aproximada (b)	Sig. aproximada
Nominal por nominal	Phi	,418			,000
	V de Cramer	,418			,000
	Coeficiente de contingencia	,386			,000
Ordinal por ordinal	Tau-b de Kendall	,147	,066	2,236	,025
	Tau-c de Kendall	,183	,082	2,236	,025
	Gamma	,230	,102	2,236	,025
	Correlación de Spearman	,166	,074	2,243	,026(c)
Intervalo por intervalo	R de Pearson	,084	,075	1,118	,265(c)
N de casos válidos		180			

a Asumiendo la hipótesis alternativa.
b Empleando el error típico asintótico basado en la hipótesis nula.
c Basada en la aproximación normal.

Tabla 9.16. Cruce de tabla con todos los estadísticos disponibles.

grupos: de 18 a 29 años, de 30 a 44, de 45 a 64, y 65 y más años[131]). ¿Y el número de amigos con el hábitat o tamaño del municipio donde se reside (e36)?

Al final de la sección 3 se interpretó una relación entre el nivel de equipamiento y la clase social. ¿Qué tipos de variables han sido utilizadas en esa relación? ¿Está bien realizado el ejercicio? Proponer la solución correcta.

6. Anexo 1: Introducción al cálculo de pares

Las medidas utilizadas para analizar la relación entre variables ordinales se fundamentan en el cálculo de pares que se expone en el presente apartado. Nosotros hemos tomado esta explicación de la obra de García Ferrando (1985), entre las páginas 245-250. El número total de pares, como se expone más abajo, es el tamaño de la muestra por el tamaño de la muestra menos uno, dividido entre dos; de modo que la tabla utilizada en el apartado 9.5 cuenta con 6.216 pares. El total de pares se divide en cinco tipos de pares: semejantes o concordantes (aquellos que se distribuyen en idéntico orden en ambas variables); desemejantes o discordantes (ordenados en orden opuesto); empatados en la variable independiente X (y no en la variable dependiente Y); empatados solo en la variable dependiente Y (y no en la variable independiente X); y pares empatados en ambas variables.

Tabla de contingencia (v08) Libros leídos en los últimos cinco meses no relacionados con los estudios * (v06) Libros relacionados con tus estudios leídos en los últimos cinco meses

		(v06) Libros relacionados con tus estudios leídos en los últimos cinco meses			
		Ninguno	**Uno**	**Dos y tres**	**Total**
(v08) Libros leídos NO relacionados...	Ninguno	**22**	10	6	38
	Uno	2	**14**	16	32
	Dos y tres	12	18	**12**	42
Total	36	42	34	112	

131. Respecto al criterio de agrupación de la edad, se ha realizado siguiendo la distribución realizada por Amando de Miguel (1997: 59) en sus estudios sobre la sociedad española. Obsérvese que se trata de la misma división por grupos de edad que se propuso al final de la sección 4 del capítulo VIII.

El cálculo de pares, siguiendo a García Ferrando (1985: 246), debe comenzar con la elección de la diagonal que une las celdillas que contienen los valores *alto-alto* y *bajo-bajo* en ambas variables, denominada por García Ferrando como *diagonal positiva*. En la tabla sobre este párrafo se aprecia ésta es la diagonal que une el extremo superior izquierdo con el extremo inferior derecho, la que presenta los valores 22–14–12, valores que han sido colocados en negrilla. Definida la diagonal positiva, procedemos con el cálculo de pares, tomado de la obra de García Ferrando (1985: 245-250).

a) *Número total de pares:*

$$\text{Pares} = \frac{N\,(N-1)}{2} = \frac{112 * 111}{2} = 6.216$$

b) *Pares semejantes o concordantes:* $\mathbf{N_s}$

22 * (14 + 16 + 18 + 12) = 1.320
10 * (16 + 12) = 280
2 * (18 + 12) = 60
14 * 12= 168
Total: 1.828

c) *Pares desemejantes o discordantes:* $\mathbf{N_d}$

12 * (10 + 6 + 14 + 16) = 552
18 * (16 + 6) = 396
2 * (10 + 6) = 32
14 * 6= 84
Total: 1.064

d) *Pares empatados solo en la variable independiente X:* $\mathbf{T_x}$

22 * (2 + 12) = 308
2 * 12 = 24
10 * (14 + 18) = 320
14 * 18= 252
6 * (16 + 12) = 168
16 * 12 = 192
Total: 1.264

e) *Pares empatados solo en la variable dependiente Y:* $\mathbf{T_y}$

22 * (10 + 6) = 352
10 * 6 = 60
2 * (14 + 16) = 60

14 * 16= 224
12 * (18 + 12) = 360
16 * 12 = 216
Total: 1.272

f) *Pares empatados simultáneamente en X e Y:* $\mathbf{T_{xy}}$
Se trata de aplicar la fórmula "f (f – 1) / 2" a cada celdilla, donde f es la frecuencia de cada celdilla

22 (22 – 1) / 2 = 231
6 (6 – 1) / 2 = 15
14 (14 – 1) / 2 = 91
12 (12 – 1) / 2 = 66
12 (12 – 1) / 2 = 66
10 (10 – 1) / 2 = 45
2 (2 – 1) / 2 = 1
16 (16 – 1) / 2 = 120
18 (18 – 1) / 2 = 153
Total: 788

7. Anexo 2: Lenguaje de sintaxis de los análisis realizados

En el apartado 7 del capítulo VIII se explicó el origen de cada uno de estos mandatos (pulsando el botón Pegar en el cuadro de diálogo correspondiente), así como el proceso de *ejecución* de cada uno.

Apartado 2: Elaboración de tabla de contingencia con dos variables

```
FREQUENCIES
    VARIABLES=v01
    /ORDER= ANALYSIS.

RECODE v01 (4=14) (7=14) (8=14) (9=14) (10=14) (15 thru 18=14) (98=99) into
v01bis.

FREQUENCIES
    VARIABLES=v01bis
    /ORDER= ANALYSIS.

CROSSTABS
    /TABLES=v01bis BY v49
```

```
/FORMAT= AVALUE TABLES
/CELLS= COUNT
/COUNT ROUND CELL.
```

Apartado 3.1: Relación entre variables nominales utilizando el Chi-cuadrado

```
CROSSTABS
  /TABLES=v01bis BY v49
  /FORMAT= AVALUE TABLES
  /STATISTIC=CHISQ
  /CELLS= COUNT
  /COUNT ROUND CELL.

CROSSTABS
  /TABLES=v01bis BY v49
  /FORMAT= AVALUE TABLES
  /STATISTIC=CHISQ
  /CELLS= COUNT EXPECTED
  /COUNT ROUND CELL.

CROSSTABS
  /TABLES=v01bis BY TITULAC
  /FORMAT= AVALUE TABLES
  /STATISTIC=CHISQ
  /CELLS= COUNT EXPECTED
  /COUNT ROUND CELL.
```

Apartado 3.2: Consideraciones a tener en cuenta en la utilización del Chi-cuadrado

```
CROSSTABS
  /TABLES=v01 BY v49
  /FORMAT= AVALUE TABLES
  /STATISTIC=CHISQ CC PHI
  /CELLS= COUNT
```

```
    /COUNT ROUND CELL.

CROSSTABS
    /TABLES=v028 BY v49
    /FORMAT= AVALUE TABLES
    /STATISTIC=CHISQ CC PHI
    /CELLS= COUNT
    /COUNT ROUND CELL.
```

Apartado 3.3: Estadísticos basados en el Chi-cuadrado

```
CROSSTABS
    /TABLES=v01bis BY v49
    /FORMAT= AVALUE TABLES
    /STATISTIC=CHISQ CC PHI
    /CELLS= COUNT
    /COUNT ROUND CELL.
```

Apartado 4.1: Análisis del interior de la tabla: cálculo y diferencia de porcentajes.

```
CROSSTABS
    /TABLES=v01bis BY v49
    /FORMAT= AVALUE TABLES
    /CELLS= COUNT ROW COLUMN TOTAL
    /COUNT ROUND CELL.
```

Apartado 4.2: Análisis del interior de la tabla: utilización de residuos

```
CROSSTABS
    /TABLES=v01bis BY v49
    /FORMAT= AVALUE TABLES
    /CELLS= COUNT RESID SRESID ASRESID
    /COUNT ROUND CELL.
```

```
CROSSTABS
    /TABLES=v01bis BY v49
    /FORMAT= AVALUE TABLES
    /CELLS= COUNT ASRESID
    /COUNT ROUND CELL.
```

Apartado 5: Utilización de test estadísticos para conocer la relación entre variables ordinales.

```
FREQUENCIES
    VARIABLES=v06 v08
    /ORDER= ANALYSIS.

MISSING VALUE V08 (99).

FREQUENCIES
    VARIABLES=v08
    /ORDER= ANALYSIS.

USE ALL.
COMPUTE filter_$=(v06 <= 3 & v08 < 4).
VARIABLE LABEL filter_$ 'v06 <= 3 & v08 < 4 (FILTER)'.
VALUE LABELS filter_$ 0 'No seleccionado' 1 'Seleccionado'.
FORMAT filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE.

FREQUENCIES
    VARIABLES=v06 v08
    /ORDER= ANALYSIS.

CROSSTABS
    /TABLES=v08 BY v06
    /FORMAT= AVALUE TABLES
    /STATISTIC=GAMMA D BTAU CTAU
    /CELLS= COUNT SRE
    /COUNT ROUND CELL.
```

```
CROSSTABS
    /TABLES=v01bis BY v49
    /FORMAT= AVALUE TABLES
    /STATISTIC=CHISQ CC PHI CORR GAMMA D BTAU CTAU
    /CELLS= COUNT COLUMN ASRESID
    /COUNT ROUND CELL.
```

Apartado 11: introducción al cálculo de pares

```
CROSSTABS
    /TABLES=v08 BY v06
    /FORMAT= AVALUE TABLES
    /STATISTIC=GAMMA D BTAU CTAU
    /CELLS= COUNT SRE
    /COUNT ROUND CELL.
```

Apartado 12: presentación de un ejemplo.

```
FREQUENCIES
    VARIABLES=v44
    /ORDER= ANALYSIS.
RECODE v44 (4=10) (1 thru 3=20) (5 thru 23=20) INTO TITULAC.
VARIABLE LABELS TITULAC 'Titulación (Sociología/no sociología)'.
VALUE LABELS TITULAC 10"Sociología" 20"No Sociología".
FREQUENCIES
    VARIABLES=titulac
    /ORDER= ANALYSIS.

FREQUENCIES
    VARIABLES=v18
    /ORDER= ANALYSIS.
RECODE v18 (90 thru 99=SYSMIS).
FREQUENCIES
    VARIABLES=v18
    /ORDER= ANALYSIS.
RECODE v18 (3=2).
VALUE LABELS v18 1"Sábado" 2"Domingo y otro día festivo" 4"Otro día no festivo".
```

```
FREQUENCIES
    VARIABLES=v18
    /ORDER= ANALYSIS.

CROSSTABS
    /TABLES=v18 BY TITULAC
    /FORMAT= AVALUE TABLES
    /STATISTIC=CHISQ PHI
    /CELLS= COUNT COLUMN
    /COUNT ROUND CELL.

CROSSTABS
    /TABLES=v18 BY TITULAC
    /FORMAT= AVALUE TABLES
    /STATISTIC=CHISQ PHI
    /CELLS= COUNT ASRESID
    /COUNT ROUND CELL.
```

Capítulo X

Tablas de contingencia con mas de dos variables

1. Objetivos didácticos del capítulo

La explicación de las tablas de contingencia de dos variables –relaciones bivariantes– da paso a las relaciones *múltiples*, a las tablas de contingencia donde intervienen tres y más variables. Éstas pueden participar formando tablas de dos dimensiones donde cada fila o columna está formada por una o varias variables, o tablas con más de dos dimensiones.

En el primer caso, las tablas de dos dimensiones son elaboradas uniendo variables que recogen las *respuestas múltiples* de determinadas preguntas, concretamente aquellas con más de una respuesta (preguntas multirrespuesta). El cuestionario utilizado a lo largo del libro (presentado en la sección 6 del capítulo segundo) dispone de un gran número de preguntas de este tipo: ¿cuáles son las *dos* situaciones que mejor definen tu actividad en tu tiempo libre (pregunta 3); ¿qué asignaturas proponen libros de lectura obligatoria (pregunta 7); ¿qué periféricos o dispositivos tienen los ordenadores de tu hogar (pregunta 17a); etc. Tal y como señalamos en el capítulo siete se trata de unas preguntas muy utilizadas en la investigación con encuesta, y con un tratamiento diferente al resto de preguntas del cuestionario.

Las tablas de más de dos dimensiones relacionan más de dos variables eliminando la influencia de la última; la tercera en el caso de tres dimensiones, la cuarta cuando se trata de tablas cuatridimensionales, etc. Dedicaremos el segundo apartado a la interpretación de tablas de tres dimensiones, y en el siguiente se explicarán las diferentes situaciones que genera la *última variable* de la tabla de contingencia. Realizaremos, de este modo, un ligero acercamiento a las relaciones múltiples entre variables, al análisis multivariable, al que nos referimos someramente en el capítulo II (apartado 2.5).

Como en el resto de capítulos, la explicación se ha llevado a cabo utilizando ejemplos realizados con el archivo de datos obtenido del cuestionario presentado en el segundo capítulo, sección 7 (ENCUESTAS ESTUDIANTES 2003_04.SAV).

2. Tablas de contingencia de respuestas múltiples categóricas

Comenzaremos la explicación de las tablas de contingencia con varias variables analizando las preguntas que presentan categorías de respuestas no excluyentes, aquellas donde los entrevistados pueden seleccionar varias de las alternativas posibles. Recordemos que éstas fueron definidas en el capítulo siete (aparado 7.4 y 7.5) como preguntas *multirespuesta* o *respuesta múltiple*, diferenciando entre preguntas multirespuesta categóricas y dicotómicas. La propia temática del capítulo siete, dedicado al análisis univariable, limitó la exposición de estas preguntas a la *unión* de preguntas (definición de conjuntos) y a la obtención de frecuencias. En el presente apartado daremos un paso más y relacionar las preguntas múltiples con otras variables.

Antes de proceder con el cruce de tablas debe realizarse un análisis de las frecuencias, tal y como hemos mantenido a lo largo de todo el libro. Con el fin de simplificar la exposición volveremos de nuevo a la variable *situaciones que mejor definen la actividad durante el tiempo libre*, pregunta 3, y cuyas respuestas se recogen en las variables v03 y v04. Su distribución de frecuencias fue presentada en la tabla 7.1, y ahí señalamos que para el 66,3% de los entrevistados la situación que mejor define su tiempo libre es *estar con la gente*, un 36,8% señaló *dedicarme tranquilamente a mis cosas y aficiones*, y otro 22,1% mostró su preferencia por *hacer muchas cosas*.

Una vez interpretadas las frecuencias llega el momento de proceder con el cruce de tablas. Utilizaremos, para ello, el menú *Analizar⇒Respuestas múltiples⇒Tablas de contingencia...* que da lugar al cuadro de diálogo de la figura 10.1. En la ventana superior izquierda aparecen todas las variables del cuestionario, y en la inferior izquierda la agrupación de variables que forman las preguntas multirespuesta; concretamente las variables *situaciones que mejor definen la actividad durante el tiempo libre* (pregunta 3) y *número de dispositivos en el ordenador (pregunta 17a)*; que fueron definidas en el séptimo capítulo[132]. En las ventanas centrales de la figura 10.1 se situarán las variables a *cruzar*, aquellas que formarán la tabla de contingencia.

En este caso nuestro objetivo es conocer las *situaciones que mejor definen la actividad durante el tiempo libre* diferenciando el sexo del entrevistado con el fin de descubrir si los hombres y las mujeres presentan diferencias en las situaciones que mejor definen su actividad durante el tiempo libre; continuando así con uno de los temas

132. Pudiera suceder que una vez analizadas las frecuencias de todo el cuestionario nos interesara volver a trabajar con estas variables en una sesión posterior. En este caso, antes de realizar la tabla de contingencia será necesario proceder de nuevo con la *Definición de conjuntos* mediante el menú *Analizar⇒Respuestas múltiples⇒Definir conjuntos...* Recuérdese que mientras no hay *conjuntos* (de variables) *definidos* la opción *Tablas de contingencia de respuestas múltiples* aparece inactiva.

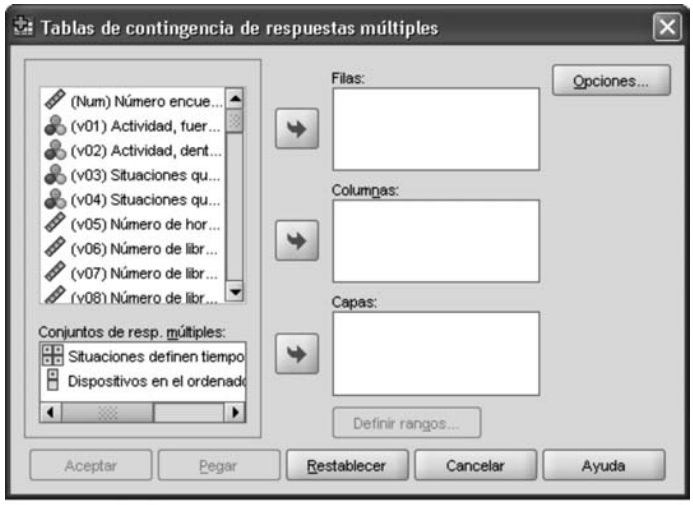


Figura 10.1. Cuadro de diálogo para realizar una tabla de contingencia con variables de respuesta múltiple (categóricas).

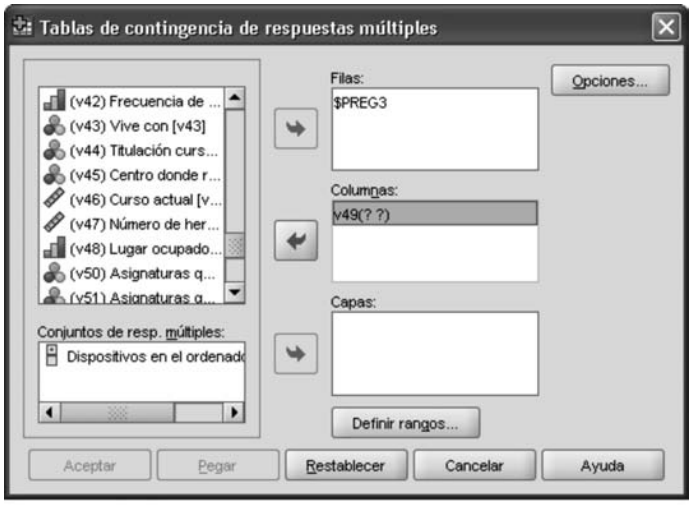


Figura 10.2. Tabla de contingencia de respuestas múltiples (categóricas) con dos variables.

planteados en el capítulo IX. Para ello será necesario desplazar *Situaciones definen...* a la ventana de las filas y el sexo del entrevistado (V49) a la ventana de las columnas (figura 10.2). En este momento el botón *Definir rangos...* cambia de color y, una vez pulsado, aparece un cuadro de diálogo para definir el valor mínimo y máximo de la variable situada en columnas (figura 10.3).

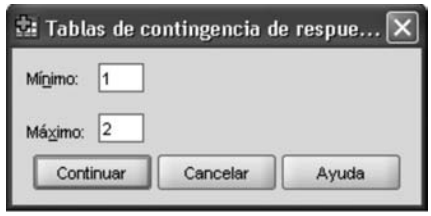


Figura 10.3. Tabla de contingencia con variables de respuesta múltiple (categóricas): definir rangos variable en columnas.

Figura 10.4. Tabla de contingencia con variables de respuesta múltiple: Opciones...

El botón *Continuar* del cuadro de diálogo de la figura 10.3 introduce estos valores en el cuadro de diálogo principal. Definida la tabla, recomendamos pulsar el botón *Opciones...* (figura 10.2) que muestra los elementos que aparecerán en el interior de las celdillas (figura 10.4). Nuestro objetivo, que recordemos consistía en conocer si los hombres y las mujeres presentan diferencias en las situaciones que mejor definen su actividad durante el tiempo libre, nos ha llevado a solicitar los porcentajes de columna. Del resto de elementos incluidos en este cuadro de diálogo la opción más importante es "*Porcentajes basados en*", que lo dejaremos como está por la explicación realizada en el capítulo VII. Lo mismo hacemos con los valores perdidos[133]. Pulsar el botón *Continuar* da paso al cuadro de diálogo principal que, tal y como se muestra en

133. Marcar la opción *Excluir los casos según lista dentro de las dicotomías/categorías* implica la eliminación de los casos que presentan valores perdidos en cualquiera de las variable que forman parte del conjunto de respuestas múltiples.

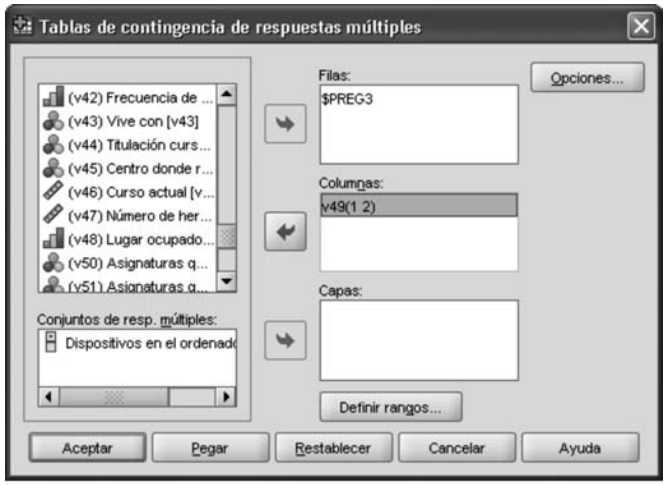


Figura 10.5. Tabla de contingencia con variables de respuesta múltiple.

la figura 10.5, aparece con el rango de la variable situada en columnas. Tras pulsar el botón *Aceptar* obtenemos los resultados que se muestran en la tabla 10.1.

En la parte superior de la tabla 10.1 se presenta, al igual que en las tablas de contingencia vistas en el capítulo anterior, el resumen de los casos; donde se indica el número de casos válidos y perdidos. 191 casos son válidos en este ejemplo. A continuación aparece la tabla de contingencia con la variable sexo en columnas y las situaciones que definen el tiempo libre en filas. Aunque en este caso la parte final de la etiqueta define con precisión el origen de esta variable ("v3+v4"), la nota (a) junto a ésta indica que se trata de una agrupación, de una variable formada por la unión de varias. Dentro de las celdillas aparece el recuento y los porcentajes verticales ("% dentro de v49"). Al pie de la tabla se indica que los porcentajes y los totales se basan en los encuestados; en los casos (tal y como se definió en la figura 10.4).

La columna de la derecha (rotulada con *total*) es la suma de las variables v03 y v04, es decir, la distribución univariante con las dos variables agrupadas que se mostró en la tabla 7.1. El análisis de las celdillas interiores debe hacerse tal y como señalamos en el capítulo anterior, en el apartado *análisis del interior de la tabla (9.4)*. Centraremos la explicación en los porcentajes, en la medida que la interpretación de los números absolutos es complicado por la diferencia entre el número de hombres y mujeres entrevistadas (71 y 120 respectivamente). Se trata de porcentajes de columna, calculados dividiendo el número de casos de la celdilla entre el número de respuestas obtenido por cada una de las categorías de la variable columna (v049)[134]. Ahora bien, a

134. (6 / 71) * 100 = 8,5; (12 / 71) * 100 = 16,9; (10 / 71) * 100 = 14,1;...

Resumen de los casos

	Casos					
	Válidos		Pérdidos		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
$PREG3*v49	191	100,0%	0	,0%	191	100,0%

Tabla de contingencia $PREG3*v49

			(v49) Genero		
			Hombre	Mujer	Total
Situaciones definen tiempo libre (v3+v4)(a)	Pasarlo bien sin hacer nada	Recuento	6	8	14
		% dentro de v49	8,5%	6,7%	7,3%
	Hacer muchas cosas	Recuento	12	30	42
		% dentro de v49	[135]*16,9%*	25,0%	22,0%
	Dedicarme a las personas más queridas	Recuento	10	26	36
		% dentro de v49	14,1%	21,7%	18,8%
	Hacer cosas de mi trabajo que tengo pendientes	Recuento	6	20	26
		% dentro de v49	*8,5%*	16,7%	13,6%
	Descansar, recuperar fuerzas	Recuento	17	24	41
		% dentro de v49	23,9%	20,0%	21,5%
	Estar con la gente, charlar, tratar a los amigos	Recuento	42	84	126
		% dentro de v49	*59,2%*	70,0%	66,0%
	Aburrirme	Recuento	6	0	6
		% dentro de v49	**8,5%**	,0%	3,1%
	Pensar, meditar	Recuento	4	4	8
		% dentro de v49	5,6%	3,3%	4,2%
	Dedicarme tranquilamente a mis cosas, aficiones, deportes	Recuento	34	36	70
		% dentro de v49	**47,9%**	*30,0%*	36,6%
Total		Recuento	71	120	191

Los porcentajes y los totales se basan en los encuestados.
a Agrupación

Tabla 10.1. Tabla de contingencia Pregunta 3 y sexo (v49); porcentajes basados en casos.

135. Se ha utilizado el subrayado para identificar aquellos porcentajes inferiores al promedio (columna *total*); mientras que la negrilla indica porcentajes notablemente superiores al promedio.

diferencia de las tablas de contingencia vistas en el capítulo anterior en este caso los porcentajes de columna no suman 100 en sentido vertical puesto que cada entrevistado ha proporcionado más de una respuesta, dos en este ejemplo[136].

Pese a este hecho, los porcentajes se interpretan exactamente igual como señalamos en el capítulo anterior: el 59,2% de los hombres señalan que la situación que mejor define su tiempo libre es *estar con la gente, charlar, tratar a los amigos*, un 47,9% de los hombres señala *dedicarme tranquilamente a mis cosas, aficiones y deportes*, el 23,9% muestra su preferencia por *descansar, recuperar fuerzas*, …, por señalar las situaciones que mejor definen el tiempo de ocio (las categorías con más elecciones, con mayores porcentajes)[137].

En el capítulo anterior indicamos que la gran riqueza del análisis de tablas de contingencia es poder comparar *transversalmente* estos porcentajes, cotejando *linealmente* las celdas de las distintas columnas. En tal caso deberíamos resaltar únicamente los porcentajes donde existen diferencias importantes entre los hombres y las mujeres, comparando cada categoría con el total. Esto sucede, en el caso de los hombres, en *dedicarme tranquilamente a mis cosas*, diferencia de 11,3 puntos; *estar con la gente, charlar con los amigos*, diferencia de –6,8 puntos, *aburrirme*, diferencia 5,4; *hacer cosas del trabajo que tengo pendientes* y *hacer muchas cosas*, diferencias de –5,1 puntos y, por último, *dedicarme a las personas más queridas*, con una diferencia de –4,7 puntos. De este modo podríamos decir que las situaciones que mejor definen el tiempo libre de los hombres son *dedicarme tranquilamente a mis cosas* y *aburrirme*; mientras que las mujeres apuestan en mayor medida por *estar con la gente, charlar con los amigos* (diferencia 4 puntos), *hacer cosas del trabajo que tengo pendientes* (diferencia 3,1 puntos) y *hacer muchas cosas* (diferencia 3 puntos). Es importante precisar, en el caso de las mujeres, que las diferencias son muy pequeñas, menores de los valores mínimos recomendados señalados en el capítulo anterior (recordar nota a pie número 118). También existen, como no, situaciones de tiempo libre que no varían según el sexo del entrevistado; esto es, que son igualmente elegidas por hombres y por mujeres. Se trata, concretamente, de pasarlo bien sin hacer nada y descansar.

Buscando *fijar* los conocimientos aprendidos en esta sección, antes de considerar nuevos contenidos, proponemos un par de ejercicios utilizando la investigación sobre *Vida Cotidiana*.

136. Como vimos en el capítulo siete los porcentajes sumarían 100 si el cálculo se hubiera realizado según el *porcentaje de respuestas*.

137. En el capítulo VII recomendamos leer un artículo de prensa para observar cómo se ha presentado una pregunta similar en un medio de comunicación. Recomendamos, de nuevo, volverlo a leer.

- Considerando los objetivos más importantes a solucionar en España[138] (pregunta 6, variables a27, a29 y a31) y en el mundo (pregunta 8, variables a39, a41 y a43), ¿presentan variación en función del sexo? ¿Y en función de la edad?
- Describe los principales motivos de discusiones entre los españoles (pregunta 29, variables b43, b45, b47 y b49). No se trata especificar *con quién* se discute *sobre qué*, sino en señalar los motivos de discusión sin diferenciar la persona con la que se discute. Estos motivos de discusión, ¿presentan variación en función de la edad? ¿Y según la zona geográfica?

Los tipos de personas –al margen de los familiares– con los que los entrevistados se relacionan habitualmente (pregunta 34, variables b63, b64 y b65[139]), ¿presentan variación en función de la edad de los entrevistados? Señala la *asociación* entre cada grupo de edad y los distintos *tipos de personas* que se presentan en la pregunta 34.

3. Tablas de contingencia de respuestas múltiples dicotómicas

El segundo tipo de preguntas multirespuesta son conocidas como dicotómicas y se diferencian de las anteriores en que consideran tan sólo una categoría de la variable. En el séptimo capítulo, apartado 7.5, se analizó la pregunta 17a que recordemos preguntaba a los entrevistados con ordenador por la presencia de una serie de dispositivos. Las respuestas eran codificadas con el valor 1 cuando se contaba con ese equipamiento, dejándose sin responder en caso contrario[140]. En la tabla 7.2 vimos que el 94,4% de los entrevistados con ordenador disponían de también de impresora, un 92,2% de lectora de CD, un 91,1% de altavoces...

138. Recuérdese la recomendación realizada en el capítulo VII, concretamente en la nota a pie número 73 (página 177), donde se aconsejaba tener cuidado con la interpretación de esta pregunta puesto que las tres variables no pueden *agregarse* de esta forma al estar ordenadas por orden de importancia. En aquel momento se señaló que para la realización del ejercicio se considerará que se tratan de tres *objetivos* igualmente importantes, no ordenados según la mayor o menor trascendencia. Esto es, olvidando que la variable a39 recoge el objetivo más importante, la a41 el segundo más importante, y la a43 el tercero más importante.
139. Considerando que se trata de tres relaciones *con igual importancia*, esto es, sin considerar el *orden* según la mayor o menor frecuencia; al igual que en el ejercicio anterior.
140. Los que no disponen de cada equipamiento no responden la variable (ver cuestionario en el apartado 6 del capítulo II) pero –por la necesidad de introducir todos los valores de los entrevistados– en la codificación del cuestionarios (sección 6 del capítulo III) recomendamos codificar esta situación con el valor 0. Ver libro de códigos de la sección 9 del capítulo III.

En este momento nos interesa conocer si el equipamiento del ordenador varía según se tenga –o no– conexión a Internet desde el hogar (pregunta 17b, v49); planteando la hipótesis que *los ordenadores conectados precisarán de un mayor equipamiento*. Solicitando las frecuencias de la variable v30[141] se aprecia que 140 estudiantes disponen de conexión a Internet desde su hogar. De los 191 entrevistados, el porcentaje de estudiantes con conexión a Internet en el hogar es del 77,8%.

Figura 10.6. Tabla de contingencia de respuestas múltiples (dicotómicas) con dos variables.

Para conocer si el equipamiento del ordenador cambia según la conexión a Internet desde el hogar será necesario realizar una tabla de contingencia entre ambas variables. Utilizaremos para ello el menú *Analizar⇒Respuestas múltiples⇒Tablas de contingencia...* colocando la variable *Número de dispositivos* en filas, y el *acceso a internet desde tu hogar* (v30) en columnas. A continuación definimos el rango de esta última, que identifica la categoría *Si* con el uno y el *No* con el dos[142], obteniendo la figura 10.6. Pulsando el botón *Opciones* seleccionamos, al igual que en el ejemplo anterior, los porcentajes de columnas y dejamos *por defecto* los porcentajes basados en casos.

141. Menú *Analizar⇒Estadísticos Descriptivos⇒Frecuencias⇒*v30. En el apartado 2 no se han solicitado las frecuencias del sexo porque ya se conocen del capítulo nueve.
142. Esta variable tenía también los valores 90 y 98 (ver libro de códigos en el apartado 9 del capítuilo III). Al seleccionar el 1 y el 2 los valores superiores no son considerados.

Resumen de los casos

	Casos					
	Válidos		Pérdidos		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
$nº_dispo*v49	180	94,2%	11	5,8%	191	100,0%

Tabla de contingencia $nº_dispo*v30

			(v30) Acceso a Internet desde el hogar		
			Si	No	Total
Dispositivos en el ordenador(a)	(v22) Dispositivos en el ordenador: IMPRESORA	Recuento	132	38	170
		% dentro de v30	94,3%	95,0%	94,4%
	(v23) Dispositivos en el ordenador: MODEM	Recuento	128	6	134
		% dentro de v30	**91,4%**	*15,0%*	74,4%
	(v24) Dispositivos en el ordenador: ALTAVOCES	Recuento	136	28	164
		% dentro de v30	**97,1%**	*70,0%*	91,1%
	(v25) Dispositivos en el ordenador: WEBCAM	Recuento	36	0	36
		% dentro de v30	25,7%	,0%	20,0%
	(v26) Dispositivos en el ordenador: LECTORA CD	Recuento	134	32	166
		% dentro de v30	**95,7%**	*80,0%*	92,2%
	(v27) Dispositivos en el ordenador: GRABADORA CD	Recuento	60	8	68
		% dentro de v30	42,9%	*20,0%*	37,8%
	(v28) Dispositivos en el ordenador: LECTORA DVD	Recuento	90	14	104
		% dentro de v30	**64,3%**	*35,0%*	57,8%
	(v29) Dispositivos en el ordenador: GRABADORA DVD	Recuento	24	0	24
		% dentro de v30	17,1%	,0%	13,3%
Total		Recuento	140	40	180

Los porcentajes y los totales se basan en los encuestados.
a Agrupación de dicotomías. Tabulado el valor 1.

Tabla 10.2. Tabla de contingencia "Número de Dispositivos" (Pregunta 17a) y conexión a internet en el hogar (v30); porcentajes basados en casos.

Los resultados se muestran en la tabla 10.2, donde se aprecia que de los 191 casos analizados el 94,2% han sido considerados como válidos, mientras que 11 –que suponen un 5,8% de los entrevistados– son definidos como perdidos. Se trata de 8 personas que no cuentan con ordenador (y que por lo tanto no se les ha preguntado si tienen o no conexión) y otras 3 que no la han respondido. La nota (a) junto al nombre de la variable en filas indica que se trata de una agrupación de dicotomías (pregunta multirespuesta dicotómica) realizada considerando el valor 1. Al pie de la tabla puede apreciarse también que los porcentajes se basan en los casos, en los encuestados.

La interpretación del interior de la tabla desvela, en primer lugar, que los entrevistados que disponen de conexión a internet en su hogar se caracterizan por tener ordenadores con más equipamientos, destacando fundamentalmente la presencia de modem (diferencia respecto al total de 17 puntos), altavoces (diferencia de 6 puntos), y lector de DVD (diferencia de 6,5 puntos). En segundo lugar puede apreciarse que la práctica totalidad de ordenadores *conectados* disponen de altavoces (únicamente el 2,9[143]% no lo tienen), y 9 de cada diez de modem. Resulta sorprendente, por otro lado, que seis personas que no disponen de conexión a internet en el hogar cuenten con ordenador con modem; lo que puede llevarnos a dudar si se ha comprendido bien la pregunta. Un análisis exhaustivo de esta pregunta –que no detallaremos aquí por motivos de espacio y por alejarse de nuestros propósitos– nos ha llevado a la conclusión que se trata de ordenadores con modem incorporado. Es importante señalar también que ninguno de los entrevistados *no conectados* dispone de WebCam y grabadora de DVD; lo cual nos lleva a pensar que podría tratarse de ordenadores con más antigüedad que los conectados a Internet. La impresora, sin embargo, se distribuye por igual entre los que están conectados y no conectados a Internet.

Se trata, como hemos visto, de una interpretación muy similar a la realizada con las preguntas multirespuesta categóricas. Con el fin de añadir una mayor complejidad a la explicación –y así llegar a comprender mejor la diferencia entre los porcentajes basados en casos y en respuestas– repetiremos el análisis anterior solicitando *Porcentajes basados en respuestas*, modificando una opción del cuadro de diálogo *Opciones* (figura 10.7).

Los resultados se muestran en la tabla 10.3; si bien se ha eliminado el resumen de los casos puesto que es el mismo que el presentado en la tabla 10.2. Obsérvese que al pie de la tabla se indica que los porcentajes y totales se basan en las respuestas. Aparte del cambio en la magnitud de los coeficientes, que se han reducido notablemente y suman 100 en sentido vertical, se aprecia también algunas diferencias cuando se comparan estos resultados con la tabla 10.2. Resulta sorprendente, en primer lugar,

143. 100 – 97,1 = 2,9

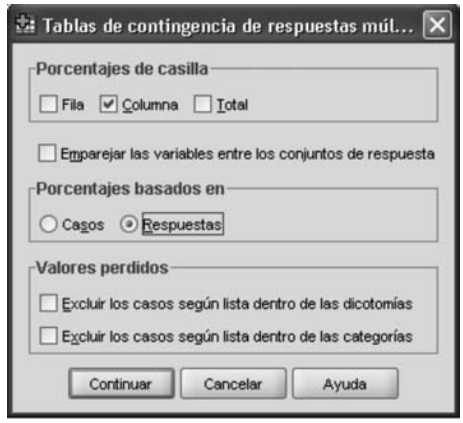


Figura 10.7. Tabla de contingencia con variables de respuesta múltiple, opciones: "Porcentajes basados en Respuestas".

la magnitud de los que poseen impresora, un 12% más entre los que no tienen acceso a Internet desde el hogar (recordemos que en la tabla 10.2 la diferencia era –tan sólo– de un 0,7%). Otro aspecto interesante es la posesión de altavoces, que en esta tabla es ligeramente superior entre los que no están conectados a Internet; justo lo contrario que desvelaba la tabla anterior. Esta misma situación se produce con el lector de CD. A la hora de explicar el motivo de esta situación tendríamos que destacar el diferente cociente en ambas tablas: la tabla 10.2 divide las frecuencias observadas de cada celdilla entre el total de entrevistados (140 con Internet y 40 sin este equipamiento); mientras que en la tabla 10.3 las frecuencias observadas son divididas entre el total de respuestas, que ascienden a 740 en el caso de los conectados a Internet y a 126 en los no conectados[144].

Esta diferencia está originada por el *objeto* interpretado; por la información que contiene la tabla. Está claro que la tabla 10.2 (basada en casos) se refiere al equipamiento presente en los ordenadores de los entrevistados. Ahora bien, ¿qué información presta la tabla 10.3?, basada en respuestas. En realidad lo que está mostrando es *como se reparte el equipamiento* en la población analizada. Esta tabla indica que –en total– se cuenta de 866 dispositivos en el ordenador (respuestas), y de éstos 740 se encuentran en los ordenadores conectados a Internet desde el hogar y 126 en los no conectados. Del total de dispositivos en los ordenadores *conectados* los más extendidos son

144. La gran diferencia entre los cocientes utilizados en una y otra tabla *favorece* a los valores más pequeños, puesto que en la primera tabla la relación tiene/no tiene Internet es de 3 a 1 (exactamente 3,15; 140 / 40); mientras que la segunda presenta una relación de 5,9 a 1 (740 / 126 = 5,87).

los altavoces (18,4%), lector de CD (18,1%) y modem (17,3%). Entre los ordenadores no conectados destaca la presencia de impresoras (30,2%), lectoras de CD (25,4%) y altavoces (25,4%).

Tabla de contingencia $nº_dispo*v30

			(v30) Acceso a Internet desde el hogar		
			Si	**No**	**Total**
Dispositivos en el ordenador(a)	((v22) Dispositivos en el ordenador: IMPRESORA	Recuento	132	38	170
		% dentro de v30	17,8%	**30,2%**	19,6%
	(v23) Dispositivos en el ordenador: MODEM	Recuento	128	6	134
		% dentro de v30	**17,3%**	4,8%	15,5%
	(v24) Dispositivos en el ordenador: ALTAVOCES	Recuento	136	28	164
		% dentro de v30	18,4%	22,2%	18,9%
	(v25) Dispositivos en el ordenador: WEBCAM	Recuento	36	0	36
		% dentro de v30	4,9%	,0%	4,2%
	(v26) Dispositivos en el ordenador: LECTORA CD	Recuento	134	32	166
		% dentro de v30	18,1%	**25,4%**	19,2%
	(v27) Dispositivos en el ordenador: GRABADORA CD	Recuento	60	8	68
		% dentro de v30	8,1%	6,3%	7,9%
	(v28) Dispositivos en el ordenador: LECTORA DVD	Recuento	90	14	104
		% dentro de v30	12,2%	11,1%	12,0%
	(v29) Dispositivos en el ordenador: GRABADORA DVD	Recuento	24	0	24
		% dentro de v30	3,2%	,0%	2,8%
Total		Recuento	740	126	866

Los porcentajes y los totales se basan en las respuestas.
a Agrupación de dicotomías. Tabulado el valor 1.

Tabla 10.3. Tabla de contingencia "Número de Dispositivos" (Pregunta 17a) y conexión a Internet en el hogar (v30); porcentajes basados en respuestas.

Terminaremos la explicación con una reflexión sobre las limitaciones de las preguntas de respuesta múltiple. Además de las señaladas al final del apartado 7.5[145], estas variables presentan el inconveniente que no es posible recodificar la variable resultante, sino que debe hacerse en las variables originales. Es decir, hay que construir la respuesta múltiple, analizar las frecuencias y, en caso que sea preciso, recodificar las variables originales para elaborar de nuevo los *conjuntos de respuestas múltiples*.

El lector habrá apreciado otra de las diferencias existentes entre éstas y las tablas de contingencia mostradas en el capítulo anterior. En las tablas multirespuesta no hay *estadísticos* que indiquen si la relación entre variables es significativa[146], estadísticos que suponen un adecuado *instrumento* en la redacción del informe de investigación cuando se están manejando una gran cantidad de tablas. En las tablas de respuesta múltiple es necesario revisarlas una a una para *desechar* (o no considerar) aquellas con relación no significativa entre variables. Además, tampoco es posible utilizar los residuos para conocer las celdillas donde se producen una mayor relación entre variables. Estos motivos, junto con los indicados en el séptimo capítulo, nos llevan a recomendar una utilización *prudente* de este tipo de preguntas.

Fijar los conocimientos aprendidos es el fin de toda actividad docente. Con el fin de facilitar esta tarea proponemos cuatro ejercicios a realizar con el archivo de la investigación sobre Vida Cotidiana de la Fundación CIRES:

- El equipamiento disponible en el hogar, ¿varían según la comunidad autónoma de residencia? ¿Y con el tamaño del municipio?
- Considerando la pregunta sobre la frecuencia con la que se comen determinados productos alimenticios (pregunta 56, variables de la c58 a la c64), y teniendo en cuenta únicamente aquellos entrevistados que *comen todos o casi todos los días* (opción 1), ¿existe variación según la clase social de pertenencia? ¿Y considerando la relación con el cabeza de familia? ¿Y respecto al tamaño del hogar (número de miembros)?

145. Necesidad de definir los conjuntos de respuestas múltiples cada vez que comienza una nueva sesión de trabajo, y una interpretación más complicada de los resultados

146. No vendrá mal recordar uno de los requisitos citados por Reynolds (1984: 19) para utilizar el Chi–Cuadrado, empleado para variables nominales como las utilizadas en los ejemplos: las categorías de las variables deben ser exhaustivas y mutuamente excluyentes.

4. Relaciones múltiples con tablas de más de dos variables. Introducción al análisis multivariable

En el capítulo nueve, concretamente en el apartado 9.3, propusimos analizar si existe relación significativa entre el grado de felicidad y el estado civil. Los resultados obtenidos se encuentran en los *materiales complementarios* (web) del capítulo 9, y su análisis muestra una relación significativa: valor V de Cramer de 0,171; con una significación de 0,000. La interpretación del interior de la tabla considerando los porcentajes de columna desvela porcentajes superiores en la categoría *muy feliz* por parte de los solteros, los casados y los que conviven en pareja. Es importante señalar el bajo porcentaje en esta categoría de los que *han estado casados*[147]. Los que conviven en pareja destacan también por sus porcentajes –más elevados que el promedio– en *bastante feliz*. Por último los que han estado casados (viudos, separados y divorciados) superan ampliamente el porcentaje promedio (total) en la categoría *Poco y nada feliz*.

Algunos lectores estarán dudando de la adecuación de las agrupación realizada en la categoría *han estado casados* en la medida que se han unido personas con perfiles muy diferentes[148]. De acuerdo que los separados, divorciados y viudos *han estado casados* anteriormente, pero se trata de situaciones muy diferentes. De momento los últimos (viudos) no han decidido *dejar* de estar casados, sino que el fallecimiento de su cónyuge les ha llevado a esa situación. En relación a los separados y divorciados debemos considerar la fecha de realización del estudio (año 1993), y la legislación en la materia, que implica que para obtener el divorcio era necesario *certificar* una separación; por lo que el estar separado era una condición previa al divorcio. La fecha de realización del estudio explica también el bajo número de personas que se encuentran en esta situación, 30 entrevistados que suponen un 2,6% de los entrevistados.

Otra posibilidad de agrupación vendría de unir los separados y divorciados con los solteros, en la medida que todos son "no casados". Ahora bien, debe tenerse en cuenta que se trata de grupos totalmente diferentes: el estilo de vida de los separados

147. Esta categoría se ha creado por la unión de los separados (22 entrevistados, un 1,9% de la muestra), divorciados (8 entrevistados, un 0,7%) y viudos (104 entrevistados, un 8,6%). En la explicación de las dos primeras magnitudes debemos tener en cuenta que se trata de un estudio cuyo trabajo de campo se realizó en el año 1993.

148. No vendrá mal considerar aquí la edad media de cada colectivo, obtenidos utilizado la segmentación del archivo explicada en el apartado 8.9:
Solteros: 28,54 años.
Casados: 47,28 años
Viviendo en pareja: 35,32 años.
Separados: 45,27 años
Divorciados: 48,20 años
Viudos: 67,81 años.

y divorciados (con hijos, pagando una hipoteca, etc.) poco tiene que ver con los solteros. Estos motivos nos llevan a considerar que lo más adecuado es eliminar a los separados y divorciados del análisis; en base a su reducido tamaño muestral (30 entrevistados, que alcanzan únicamente el 2,5% de la muestra). Debe tenerse en cuenta que lo treinta entrevistados separados y divorciados no pueden proporcionar una imagen representativa de la realidad de estos colectivos. Además, su escaso número no altera sustancialmente la muestra cuando se opta por eliminarlos.

En la tabla 10.4 se presenta la relación entre el grado de felicidad y estado civil sin los separados y divorciados, donde se aprecia su similitud con la tabla comentada anteriormente (mostrada en los *materiales complementarios* del capítulo 9). La V de Cramer presenta una magnitud de 0,170, con una significación de 0,000. Tan sólo la última columna presenta alguna diferencia, diferencias que no son importantes porque el tamaño muestral de los viudos (103 entrevistados) –frente al escaso número de separados y divorciados– ha permitido *mantener* los hallazgos detectados anteriormente. No obstante, estaremos de acuerdo que la tabla 10.4 representa la realidad mejor que la anterior.

			Estado civil			Total
			Solteros	**Casados y conviven**	**Viudos**	
Grado de Felicidad actual	Poco y nada feliz	Recuento	49	94	45	188
		% de est civil	16,3%	*12,4%*	**43,7%**	16,2%
	Bastante feliz	Recuento	209	565	52	826
		% de est civil	69,7%	**74,2%**	50,5%	71,0%
	Muy feliz	Recuento	42	102	6	150
		% de est civil	**14,0%**	**13,4%**	5,8%	12,9%
Total		Recuento	300	761	103	1164
		% de est civil	100,0%	100,0%	100,0%	100,0%

Tabla 10.4. Relación entre grado de felicidad y estado civil, porcentajes de columna.

El análisis de los residuos corregidos implicaría prestar atención a lo realmente significativo de esta tabla[149], como es la relación entre viudos y estar *poco y nada feliz*, con un residuo de 8,0 (tabla 10.5). Otras asociaciones importantes son estar viudo y NO[150]

149. Aprovechamos para *dar un repaso* a la interpretación de los residuos estandarizados corregidos con el fin de mostrar la gran ayuda que proporcionan en la interpretación de tablas de contingencia.

150. Se añade el *no* por la relación negativa del porcentaje.

			Estado civil			Total
			Solteros	Casados y conviven	Viudos	
Grado de Felicidad actual	Poco y nada feliz	Recuento	49	94	45	188
		Res corregido	,1	**-4,8**	**8,0**	
	Bastante feliz	Recuento	209	565	52	826
		Res corregido	-,6	**3,4**	**-4,8**	
	Muy feliz	Recuento	42	102	6	150
		Res corregido	,7	,7	**-2,2**	
Total		Recuento	300	761	103	1164

Tabla 10.5. Relación entre grado de felicidad y estado civil, residuos estandarizados corregidos.

ser *bastante feliz*, con un residuo de -4,8; *convivir en pareja* y estar *bastante feliz*, con un residuo de 3,4; y estar viudo y NO ser *muy feliz*.

Resulta llamativo el bajo estado de felicidad de los viudos, como si la convivencia con una persona de otro sexo fuera la *situación determinante* en el estado de la felicidad. Revisando el resto de tablas de contingencia de la variable grado de felicidad (ver *materiales complementarios* del capítulo 9) resulta llamativo la relación de esta variable con la edad, con un valor Tau-c de –0,115. Observando el interior de la tabla 10.6, y considerando el porcentaje de personas *poco y nada* feliz destaca la baja felicidad de los *talludos*[151] (20%) y más aún de los *mayores* (23,1%). Los *jóvenes* y *maduros* presentan –por otro lado– los porcentajes más bajos de *poco* y *nada feliz*. Dentro de la categoría *bastante feliz* destacan únicamente los maduros, con un 73,1%, si bien es necesario tener en cuenta que esta diferencia –de un par de puntos– no llega a ser significativa en los residuos. Respecto a los *muy felices* son reseñables los elevados porcentajes de los más jóvenes (18,4%) y algo más bajos en los *talludos* y *mayores* (interpretación contraria a la primera fila).

Reflexionando sobre los resultados presentados en ambas tablas resulta sorprendente el bajo grado de felicidad de los viudos y las personas de más edad. Otro aspecto llamativo es la coincidencia –en la categoría *bastante feliz*– de los que conviven en pareja y los que tienen entre 30 y 64 años; si bien la mayor relación se produce en el grupo entre 30 y 44 años. Ambas situaciones nos inducen a sospechar de la verdadera relación entre el estado civil y la felicidad. Ante la sospecha de contar con una relación no *directa*, una relación *intervenida* o *provocada* por una variable extraña a la relación considerada (variable *interviniente* que actúa como *perturbadora*) hemos llevado

151. Hemos adoptado aquí la definición de grupos de edad realizada por De Miguel (1997: 59) en sus estudios sobre la realidad española.

			Edad				Total
			De 18 a 29 años (jóvenes)	**De 30 a 44 años (maduros)**	**De 45 a 64 años (talludos)**	**65 y más (mayores)**	
Felicidad actual	Poco y nada feliz	Recuento	40	40	73	45	198
		% de Edad	*12,9%*	*12,5%*	**19,7%**	**23,1%**	16,6%
		Residuos corregidos	*-2,0*	*-2,3*	**2,0**	**2,7**	
	Bastante feliz	Recuento	212	234	266	134	846
		% de Edad	68,6%	**73,1%**	71,9%	68,7%	70,9%
		Residuos corregidos	-1,0	1,0	,5	-,7	
	Muy feliz	Recuento	57	46	31	16	150
		% de Edad	**18,4%**	14,4%	*8,4%*	*8,2%*	12,6%
		Residuos corregidos	**3,6**	1,1	*-2,9*	*-2,0*	
Total		Recuento	309	320	370	195	1194
		% de Edad	100,0%	100,0%	100,0%	100,0%	100,0%

Tabla 10.6. Relación entre grado de felicidad y edad. Porcentajes de columna y residuos estandarizados corregidos.

a cabo un análisis entre el grado de felicidad y el estado civil eliminando la influencia de la edad. Para ello se ha utilizado las tablas de contingencia vistas en el capítulo anterior, mediante el menú *Analizar⇒Estadísticos descriptivos⇒Tablas de contingencia*. En el cuadro de diálogo resultante será introducido el grado de felicidad en filas, el estado civil en columnas, y la edad categorizada en la tercera ventana del cuadro de diálogo (figura 10.8). A continuación solicitaremos los estadísticos de asociación correspondientes, concretamente el Chi-cuadrado y la V de Cramer.

Esta orden crea una tabla que analiza –es importante insistir en ello– la relación entre el grado de felicidad y el estado civil eliminando la influencia de la edad. Es lo que se conoce como *relación parcial*; relación entre dos variables eliminando la influencia de una tercera. Al tratarse de un análisis exploratorio que genera una gran cantidad de resultados recomendamos marcar la opción suprimir tablas.

No presentaremos la totalidad de los resultados, tan sólo los niveles de significación de los estadísticos solicitados (tabla 10.7). El elevado número de celdillas con frecuencia esperada menor que cinco –que alcanza el 33,3% en los maduros y talludos– precisa un análisis detallado del interior de las tablas que permite verificar que esta situación se produce fundamentalmente por el escaso tamaño muestral de los viudos. Tras eliminar esta categoría se vuelven a efectuar los análisis, obteniendo los resulta-

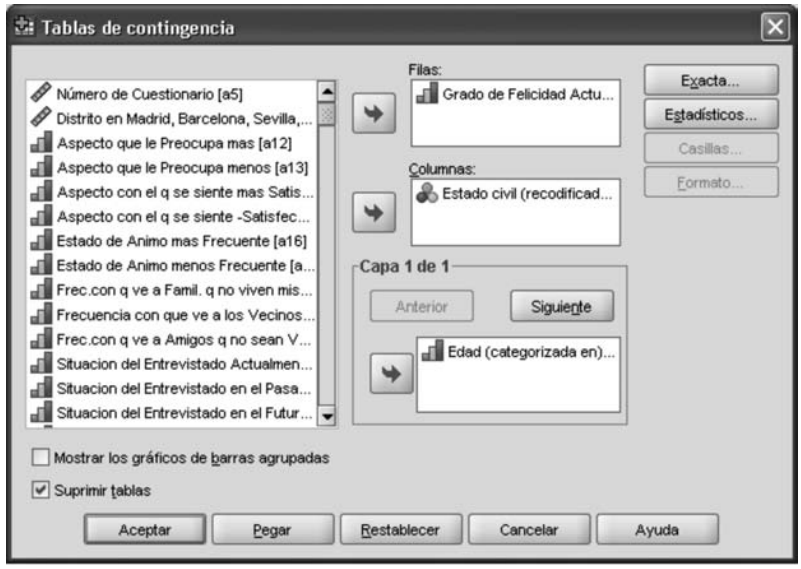


Figura 10.8. Cuadro de diálogo Tablas de contingencia con tres variables.

dos de la tabla 10.8; que muestra la desaparición de la relación entre grado de felicidad y estado civil en TODOS los grupos de edad considerados. Obsérvese que –de nuevo– hay dos grupos de edad donde el 33,3% de las celdillas muestran una frecuencia esperada menor que cinco. En el capítulo IX, apartado 3.2, señalamos que este tamaño debe ser menor que el 20%, puesto que cuando es mayor el Chi-Cuadrado tiende a aumentar *ficticiamente*. Esto se produce porque el numerador es una suma cuadrática de diferencias entre las frecuencias observadas y la teóricas. Cuando las teóricas son muy pequeñas esta suma aumenta notablemente, lo que presentará grandes tamaños del Chi-Cuadrado y –lógicamente– relaciones significativas. Por este motivo se recomienda que las frecuencias teóricas inferiores a 5 no superen el 20% de as celdillas (Reynolds, 1984: 19): En definitiva, que cuando hay muchas celdillas con frecuencia esperada menor que cinco el valor del estadístico aumenta "artificialmente".

Ahora bien, esta situación no afecta al caso que nos ocupa puesto que contamos con un valor de Chi–Cuadrado no significativo; un valor que se descendería aún más si se redujera el tamaño de las tablas. Con el fin de disminuir este gran porcentaje de celdillas con frecuencias menores a 5 procedemos a unir las categorías "Bastante feliz" y "muy feliz" en el colectivo mayor de 45 años[152]; formando así una tabla de 2 x 2

152. Se trata del procedimiento *Recodificar en distintas variables... Si los casos*; mostrado en la figura 8.11, página 207.

Pruebas de chi-cuadrado

Edad (caracterizada en cuatro categorías)		Valor	gl	Sig. asintótica (bilateral)
De 18 a 29 años (jóvenes)	Chi-cuadrado de Pearson	2,483(a)	2	,289
	Razón de verosimilitudes	2,484	2	,289
	Asociación lineal por lineal	2,467	1	,116
	N de casos válidos	304		
De 30 a 44 años (maduros)	Chi-cuadrado de Pearson	6,414(b)	4	,170
	Razón de verosimilitudes	5,952	1	,203
	Asociación lineal por lineal	4,002	4	,045
	N de casos válidos	311		
De 45 a 64 años (talludos)	Chi-cuadrado de Pearson	24,868(c)	4	,000
	Razón de verosimilitudes	20,177	1	,000
	Asociación lineal por lineal	5,920	4	,015
	N de casos válidos	355		
65 y más (mayores)	Chi-cuadrado de Pearson	21,155(d)	4	,000
	Razón de verosimilitudes	21,976	4	,000
	Asociación lineal por lineal	7,287	1	,007
	N de casos válidos	192		

a 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 11,88.
b 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 6,45.
c 2 casillas (33,3%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 2,24.
d 2 casillas (33,3%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 1,14.

Medidas simétricas

Edad (caracterizada en cuatro categorías)			Valor	Sig. aproximada
De 18 a 29 años (jóvenes)	Nominal por nominal	Phi	,090	,289
		V de Cramer	,090	,289
	N de casos válidos		304	
De 30 a 44 años (maduros)	Nominal por nominal	Phi	,144	,170
		V de Cramer	,102	,170
	N de casos válidos		311	
De 45 a 64 años (talludos)	Nominal por nominal	Phi	,265	,000
		V de Cramer	,187	,000
	N de casos válidos		355	
65 y más (mayores)	Nominal por nominal	Phi	,332	,000
		V de Cramer	,235	,000
	N de casos válidos		192	

a Asumiendo la hipótesis alternativa.
b Empleando el error típico asintótico basado en la hipótesis nula.

Tabla 10.7. Estadísticos obtenidos de la relación grado de felicidad y estado civil, eliminado la influencia de la edad.

Pruebas de chi-cuadrado

Edad (caracterizada en cuatro categorías)		Valor	gl	Sig. asintótica (bilateral)
De 18 a 29 años (jóvenes)	Chi-cuadrado de Pearson	2,483(a)	2	,289
	Razón de verosimilitudes	2,484	2	,289
	Asociación lineal por lineal	2,467	1	,116
	N de casos válidos	304		
De 30 a 44 años (maduros)	Chi-cuadrado de Pearson	5,279(b)	2	,071
	Razón de verosimilitudes	4,906	2	,086
	Asociación lineal por lineal	4,693	1	,030
	N de casos válidos	307		
De 45 a 64 años (talludos)	Chi-cuadrado de Pearson	1,056(c)	2	,590
	Razón de verosimilitudes	1,212	2	,546
	Asociación lineal por lineal	,945	1	,331
	N de casos válidos	324		
65 y más (mayores)	Chi-cuadrado de Pearson	5,509(d)	2	,064
	Razón de verosimilitudes	5,573	2	,062
	Asociación lineal por lineal	5,086	1	,024
	N de casos válidos	125		

a 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 11,88.
b 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 6,45.
c 2 casillas (33,3%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 2,24.
d 2 casillas (33,3%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 1,14.

Medidas simétricas

Edad (caracterizada en cuatro categorías)			Valor	Sig. aproximada
De 18 a 29 años (jóvenes)	Nominal por nominal	Phi	,090	,289
		V de Cramer	,090	,289
	N de casos válidos		304	
De 30 a 44 años (maduros)	Nominal por nominal	Phi	,131	,071
		V de Cramer	,131	,071
	N de casos válidos		307	
De 45 a 64 años (talludos)	Nominal por nominal	Phi	,057	,590
		V de Cramer	,057	,590
	N de casos válidos		324	
65 y más (mayores)	Nominal por nominal	Phi	,210	,064
		V de Cramer	,210	,064
	N de casos válidos		125	

a Asumiendo la hipótesis alternativa.
b Empleando el error típico asintótico basado en la hipótesis nula.

Tabla 10.8. Estadísticos obtenidos de la relación grado de felicidad y estado civil, eliminado la influencia de la edad. Análisis limitado a los solteros y convivientes en pareja (eliminados los que han estado casados).

donde las celdas con frecuencias esperadas (teóricas) inferiores a 5 se han reducido al 25% en ambas tablas, descendiendo los valores del Chi-Cuadrado al 0,38 en la tabla de 45 a 64 años (en la tabla 10.7 era de 1,056) y al 4,798 (anteriormente era de 5,509). Esta misma explicación sirve para interpretar las diferencias entre los valores Chi–Cuadrado de las tablas 10.7 y 10.8: las reducciones en esta última se producen por la disminución del número de celdillas con frecuencias teóricas menores que 5.

Para poder concluir si la variable neutralizada (la edad) influye más que la otra variable independiente (el estado civil) Reynolds (1984: 76-78) recomienda elaborar un *Indice de relación Parcial* mediante la suma ponderada de cada uno de los valores de la tabla. Si este índice está cercano a cero indicará que la relación entre las variables X e Y es muy débil cuando se elimina el influjo de la tercera. Si, por el contrario, el valor de este índice es muy alto interpretaremos que la tercera variable apenas influye. Como puede apreciarse en el cuadro 10.1 el índice de relación parcial obtenido es de 0,1059, notablemente menor que el obtenido en la relación directa entre grado de felicidad y estado civil (que recordemos era de 0,170); lo que implica que la edad influye en el grado de felicidad más que el estado civil.

$$IRP = \frac{(0{,}090 * 304) + (0{,}131 * 307) + (0{,}057 * 324) + (0{,}210 * 125)}{1.060} = 0{,}1059$$

Cuadro 10.1. índice de relación parcial, con los estadísticos de la tabla 10.7.

Sintetizando, al eliminar la influencia de la edad desaparece la relación entre grado de felicidad y estado civil; lo que implica –necesariamente– una elevada relación entre estado civil y edad. Esta situación se aprecia con precisión en la tabla 10.9, puesto que el 70% de los solteros tienen menos de 29 años (en la nota a pie número 148 vimos que la media de edad de los solteros es de 28,54 años); uno de cada tres casados tienen entre 30 y 44 años, y un 39% entre 45 y 54 años. El 64,4% de los viudos tiene más de 64 años, y un 31,7% entre 45 y 64[153]. De hecho, el valor del estadístico V de Cramer entre ambas variables asciende a 0,502; el mayor de todos los considerados a lo largo del libro. En numerosas ocasiones no aparecen relaciones tan claras, puesto que –normalmente– suele existir relación en algunos de los grupos de la variable *interviniente* (la edad en este caso) y no en otros.

153. La edad media de los casados y de los que viven en pareja es de 47,55 años, edad que aumenta hasta el 67,8 en el caso de los viudos (recordar nota a pie número 148, página 303).

Es en este momento –una vez que se ha comprobado la exactitud de todas las relaciones detectadas– cuando se procede con la elaboración de gráficos que ilustren el informe. Puede emplearse la opción *mostrar gráficos de barras agrupadas* del menú tablas de contingencia (parte inferior de la figura 10.8) o el menú *Gráficos* mostrado en la figura 4.9. Nuestra recomendación es emplear este último por el mayor número de gráficos disponibles (*Gráfico de Barras⇒Agrupado ó Apilado*; *Gráfico de Líneas⇒Múltiple*). Aunque cuando planificamos este trabajo pensábamos en dedicar un capítulo a los gráficos, las dimensiones actuales del libro –más largo de lo que planteamos en un primer momento–, unido al gran número de trabajos monográficos sobre el tema y a su excelente calidad, nos ha llevado a desistir de este empeño. Nos limitaremos a recomendar alguno, por ejemplo el elaborado por Sevilla Moróder (2006) y el –ya clásico– texto de Alaminos (1993).

			Estado civil			Total
			Soltero	Casado y convive	Viudo	
Edad	De 18 a 29 años (jóvenes)	Recuento	211	95	0	306
		% de Estado civil	70,1%	12,4%	,0%	26,2%
	De 30 a 44 años (maduros)	Recuento	55	253	4	312
		% de Estado civil	18,3%	33,1%	3,8%	26,7%
	De 45 a 64 años (talludos)	Recuento	25	300	33	358
		% de Estado civil	8,3%	39,3%	31,7%	30,6%
	65 y más (mayores)	Recuento	10	116	67	193
		% de Estado civil	3,3%	15,2%	64,4%	16,5%
Total		Recuento	301	764	104	1169
		% de Estado civil	100,0%	100,0%	100,0%	100,0%

Tabla 10.9. Relación entre estado civil y edad.

Señalar, por último, una limitación de esta técnica, como es la necesidad de contar con grandes tamaños muestrales con el fin de que los subgrupos creados por las distintas variables tengan tamaños adecuados para poder calcular el Chi–Cuadrado o los estadísticos correspondientes. Por este motivo se ha utilizado en la explicación el archivo de la investigación sobre vida cotidiana, basado en una muestra de 1.200 casos, en vez del archivo empleado a lo largo del libro, que recoge la información de 191 casos.

Buscando *fijar* los conocimientos aprendidos en esta sección, antes de considerar nuevos contenidos, proponemos un par de ejercicios utilizando la investigación

sobre *Vida Cotidiana*. El primero busca detectar si existe relación entre el nivel de estudios (e21) y el estado civil (e12) y, en caso de que así sea, comprobar si las personas con menos estudios *se casan más* que los tienen estudios elevados.

El segundo ejercicio se centra en la pregunta 17 (b3), referida a la propiedad de la vivienda: ¿La vivienda donde usted reside es de su propiedad, de su familia, o es alquilada? Se plantea como hipótesis a comprobar que las personas con bajos estudios (sin estudios y estudios primarios) presentan una gran propiedad en la vivienda (viviendas propias y pagadas); mientras que los colectivos con mayores estudios se caracterizan por no disponer de vivienda en propiedad.

5. Teoría sobre relaciones múltiples[154]

Finalizaremos el capítulo con una reflexión teórica –que surge del ejemplo expuesto en el apartado anterior– con el fin de *explicitar* las distintas relaciones entre variables en tablas de más de dos dimensiones. Respecto a la influencia de otras variables en la relación entre grado de felicidad y estado civil, la experiencia de otras investigaciones recomienda adoptar una postura crítica hacia este hecho debido a las peculiaridades sociodemográficas de nuestro país, donde la mayor parte de los jóvenes están solteros y una gran parte de las personas de más edad están casadas, tal y como vimos en la tabla 10.9. Debemos dudar si la relación entre jóvenes-solteros y casados-mayores implica que la influencia del estado civil está condicionada por la edad, o son variables que influyen de forma independiente.

Esta *relación* entre el estado civil y la edad nos invita a elaborar una hipótesis de trabajo que plantea que *pese a la relación entre estado civil (X) y grado de felicidad (Y), ésta descenderá notablemente –llegando incluso a desaparecer– cuando se elimine el influjo de la edad (Z)*. Ahora bien, antes de probar la vigencia de esta hipótesis será preciso realizar una definición de conceptos donde se explique el rol que adopta la tercera variable (variable de control).

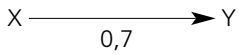
De modo que en este apartado contextualizaremos la relación entre variables realizada en el apartado anterior, explicando el interés de trabajar con tablas de contingencia formadas por tres o más variables. Se trata de análisis más complejos al introducir una tercera variable (llamémosle Z) cuya función es *controlar* la relación entre ambas variables X e Y. Lo normal es realizar varias *tablas parciales* para cada catego-

154. Hemos realizado un desarrollo más amplio de estos aspectos en Díaz de Rada, 1999: 199-221.

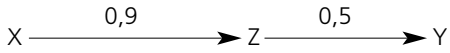
ría de Z: relación entre X e Y para Z_1, relación entre X e Y para Z_2, relación entre X e Y para Z_m. Las magnitudes de asociación entre X e Y en cada una de estas tablas parciales pueden cambiar al ser eliminado el influjo de la variable control Z[155]. Es decir, al construir una tabla (o tres tablas parciales) con las tres variables, la relación entre X e Y se convierte en una *relación parcial* al eliminar la influencia de Z, al controlar el efecto de esta última. La asociación entre variables de tablas parciales es conocida en la literatura especializada como *asociación* (o *relación*) *condicional*, mientras que la relación entre X e Y sin tener en cuenta ninguna otra variable es definida como *relación* (o *asociación*) *marginal*.

La variable que elegimos como *control*, la tercera variable de la tabla, puede adoptar diversos roles al detectar la relación entre X e Y; tal y como se muestra en el cuadro 10.2. En los siguientes párrafos describimos cada uno:

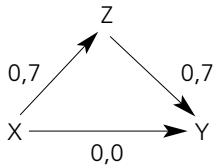
a) Relación directa:



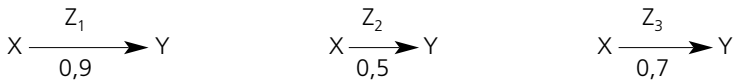
b) Variable interviniente Z:



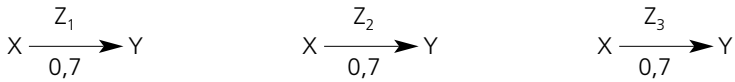
c) Relación espuria:



d) Existencia de interacciones:



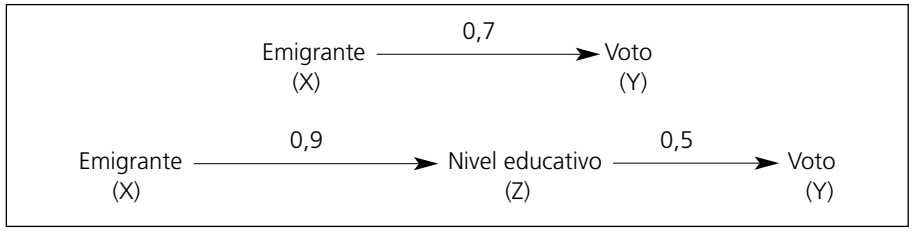
e) Ausencia de efecto de la variable Z: (independencia condicional)



Cuadro 10.2. tipos de relaciones entre tres variables.

155. Debe quedar claro que siempre es una relación entre X e Y.

1. *Variable interviniente* para explicar fenómenos. La tercera variable (Z) adopta este nombre cuando es utilizada para explicar la relación entre la variable independiente y la dependiente.
   Un ejemplo ayudará a comprender este concepto: supongamos una investigación sobre voto político donde un investigador ha encontrado una alta relación entre ser hijo de emigrantes (X) y el voto político (Y). Este investigador decide introducir una tercera variable (Z), el nivel educativo, a fin de comprobar si el *bajo nivel educativo* de los hijos de emigrantes influye en el voto político. Este investigador plantea la hipótesis que el ser hijo de emigrante está relacionado con el nivel educativo conseguido, pero no al revés; y por ello la educación es una variable interviniente. Si las personas con el mismo nivel educativo (sean hijos de emigrantes o no) muestran un comportamiento electoral similar, la variable *interviniente* estará desvelando que el ser hijo de emigrante afecta al voto político no tanto por si misma, sino porque los hijos de emigrantes alcanzan un determinado nivel educativo.



Cuadro 10.3. ejemplo de variable interviniente.

Otro ejemplo muy ilustrativo es el análisis de la relación entre las pérdidas producidas por un incendio y el número de bomberos que han acudido a apagarlo. Si alguien recoge y analiza estos datos verá que existe una elevada relación entre ambas. La dificultad para interpretar esta relación podría llevarnos a creer que son los bomberos los que producen los destrozos. Es evidente que esta interpretación es muy poco consistente, de modo que nos encontramos con una relación estadística entre variables que –a priori– carece de interpretación sociológica. Las dificultades en la interpretación le llevan al investigador a introducir una tercera variable, la magnitud del incendio, pudiendo comprobar así que la relación entre número de bomberos y pérdidas producidas por el incendio cambia notablemente. Es decir, no es que el número de bomberos genere más pérdidas en cada incendio, sino que es la intensidad del fuego la que fija el número de bomberos necesarios, y esta intensidad es la que genera los destrozos.

2. *Detección de relaciones espurias*: Responde a la pregunta de si la naturaleza y la fuerza de la asociación cambia en las tablas parciales. La relación espuria se producirá cuando al tener en cuenta la tercera variable (Z) desaparezca la relación entre la variable independiente (X) y la dependiente (Y). Dicho de otra forma, cuando la *relación parcial* entre la variable independiente y la dependiente sea cero.

   Supongamos que un investigador está analizando la relación entre la edad (X) y el voto político (Y), a fin de probar la hipótesis de que los más jóvenes cada vez presentan actitudes y comportamientos políticos más conservadores. Para ello se han recogido los datos mostrados en el tabla 10.10.a. En los jóvenes el partido conservador logra un 13,6% (56,4% – 42,8%) más de votos que en los adultos. El análisis del coeficiente Phi presenta un valor de 0,135 que es significativo al 0,000.

   La *relación marginal* mostrada en la tabla "a" permite aceptar la hipótesis planteada, si bien se trata de una relación entre voto y edad no muy alta (aunque significativa). Esta situación crea dudas en el investigador y decide incluir una tercera variable en el modelo, el nivel de estudios (tabla b). Así, en primer lugar repite esta relación para el grupo de estudios bajos (valor *Phi* 0,003, no significativo) y para el colectivo de estudios medios (valor *Phi* 0,014, no significativo). Es decir, al eliminar la influencia del nivel de estudios la relación entre la edad y el voto político desaparece, detectándose así la presencia de una relación *espuria* entre ambas.

   El nivel de estudios ha permitido detectar una relación espuria entre la edad y el voto político por lo que el investigador se plantea hasta que punto será el nivel educativo el factor que influye en el conservadurismo. Para ello elabora la tabla "c", donde se analiza la relación marginal entre el nivel de estudios y el voto, obteniendo un valor *Phi* de 0,396; el mayor de los localizados hasta el momento.

   A fin de detectar la posible influencia de la edad se vuelve a realizar este análisis por separado en los jóvenes y en los adultos (tabla "d"), encontrando valores de *Phi* altamente significativos que indican una relación de 0,341 en los jóvenes y de 0,405 en los adultos. Lo que desvela que existe relación entre los estudios y el voto político en ambos colectivos, si bien es mayor en los adultos que en los jóvenes.

   Considerando los porcentajes de la tabla se observa una diferencia de 38,9 puntos en el caso de los jóvenes (66,2% – 27,3%), y de 40,6 puntos (66,5% – 25,9%) en el nivel de conservadurismo de los adultos (los que tienen bajos estudios muestran un mayor conservadurismo); manifestándose más conservadores las personas con menos estudios. Conviene tener en cuenta que esta

Tabla "a"

		Partido Conservador	Otros	Total	% Conserv.	Phi (**)	Signif
	Jóvenes	270	209	479	56,4	0,135	0,000
	Adultos	224	299	523	42,8		

Tabla "b"

		Partido Conservador	Otros	Total	% Conserv.	Phi (*)	Signif
Estudios bajos	Jóvenes	237	121	358	66,2	0,003	0,939
	Adultos	145	73	218	66,5		
Estudios medios	Jóvenes	33	88	121	27,3	0,014	0,772
	Adultos	79	226	305	25,9		

Tabla "c"

	Partido Conservador	Otros	Total	% Conserv.	Phi (**)	Signif
Estudios bajos	382	194	572	66,3	0,396	0,000
Estudios medios	112	314	426	26,3		

Tabla "d"

		Partido Conservador	Otros	Total	% Conserv.	Phi (**)	Signif
Jóvenes	Estudios bajos	237	121	358	66,2	0,341	0,000
	Estudios medios	33	88	121	27,3		
Adultos	Estudios bajos	145	73	218	66,5	0,405	0,000
	Estudios medios	79	226	305	25,9		

(*) Relación condicional.
(**) Relación marginal

Tabla 10.10. Relación entre edad, nivel de estudios y voto político.

diferencia es mayor que la detectada al principio: 13,6%, obtenido de restar 56,4% menos 42,8%. Obsérvese, por otro lado, las escasas diferencias entre los jóvenes y adultos con estudios bajos (66,5 – 66,2), y entre los jóvenes y adultos con estudios medios (25,9 – 27,3).

3. *Localización de interacciones*: Se produce cuando la relación entre la variable dependiente (X) y la independiente (Y) cambia según las distintas categorías de la tercera variable (Z): la relación entre X e Y es fuerte en la primera categoría de la variable Z, se debilita en la segunda, vuelve a aumentar en la tercera, etc. En esta situación estaríamos en presencia de una interacción entre la tercera variable (variable control) y el resto de variables de la tabla. Dicho con otras palabras la variable control determina el tipo de relación existente entre las variables.

   Es posible también que la relación entre la variable dependiente y la independiente, tras introducir la tercera variable, genere una ligera reducción en los coeficientes de asociación de cada una de las tablas parciales, pero que esta relación continúe siendo significativa. Esta situación estaría provocada por la existencia de relaciones espurias en determinadas categorías de la variable control.

   Podemos señalar la presencia de una pequeña interacción en el ejemplo anterior (tabla 10.10) cuando analizábamos la relación entre nivel de estudios y voto político. El coeficiente *Phi* en los jóvenes es de 0,341, mientras que en los adultos es de 0,405. Aunque la magnitud de la relación es diferente, consideramos que ésta es muy pequeña para poder hablar de presencia de una interacción entre variables.

   Cuando no existen interacciones, es decir cuando la relación entre la variable X y la Y no cambian en ninguna de las categorías de Z estamos en presencia de una *asociación homogénea*.
4. *Ausencia de efecto:* Una última posibilidad es que la tercera variable no genere ninguna influencia en la relación entre la variable X y la Y, concluyendo que ésta no tiene ninguna influencia en la variable dependiente Y.

Aquí se han presentado la mayor parte de las relaciones entre variables Hemos prestado mayor importancia a la detección de relaciones espurias por su importancia en el ámbito de la investigación con encuesta.

6. Anexo 1: Lenguaje de sintaxis de los análisis realizados

Tal y como se ha procedido en los capítulos anteriores, finalizamos presentando el lenguaje de sintaxis de SPSS con los análisis realizados. Recuérdese que en el apartado 7 del capítulo VII se explicó el origen de cada uno de estos mandatos y su proceso de *ejecución*.

Apartado 2: Tablas de contingencia con variables de respuesta múltiple categóricas

```
MULT RESPONSE
    GROUPS=$preg3 'Situaciones definen el tiempo libre (v3+v4)' (v03 v04 (1,11))
    /FREQUENCIES=$preg3.

MULT RESPONSE
    GROUPS=$preg3 'Situaciones definen el tiempo libre (v3+v4)' (v03 v04 (1,11))
    /VARIABLES=v49(1 2)
    /TABLES=$preg3 BY v49
    /CELLS=COLUMN
    /BASE=CASES.
```

Apartado 3: Tablas de contingencia con variables de respuesta múltiple dicotómicas

```
MULT RESPONSE
    GROUPS=$nº_dispo 'Dispositivos en el ordenador' (v22 v23 v24 v25 v26 v27
    V28 v29 (1))
    /FREQUENCIES=$nº_disp.

FREQUENCIES
    VARIABLES=v30
    /ORDER ANALYSIS.

MULT RESPONSE
    GROUPS=$nº_dispo 'Dispositivos en el ordenador' (v22 v23 v24 v25 v26 v27
    V28 v29 (1))
    /VARIABLES=v30(1 2)
    /TABLES=$nº_dispo BY v30
    /CELLS=COLUMN
    /BASE=CASES.

MULT RESPONSE
    GROUPS=$nº_dispo 'Dispositivos en el ordenador' (v22 v23 v24 v25 v26 v27
    V28 v29 (1))
    /VARIABLES=v30(1 2)
```

```
    /TABLES=$nº_dispo BY v30
    /CELLS=COLUMN
    /BASE=RESPONSES.
```

Apartado 4: Relaciones múltiples entre varias variables. Introducción al análisis multivariable

```
FREQUENCIES
    VARIABLES=a54
    /ORDER ANALYSIS.

RECODE A54 (1=1)(2=1) (3=2) (4=3) (9=SYS) INTO A54RECO.
VARIABLE LABELS A54RECO "GRADO DE FELICIDAD ACTUAL
(RECODIFICADO EN 3 CATEGORÍAS)".
VALUE LABELS A54RECO  1'POCO Y NADA FELIZ'
                      2'BASTANTE FELIZ'
                      3'MUY FELIZ'.
FREQUENCIES
    VARIABLES=a54 a54reco
    /ORDER ANALYSIS.

FREQUENCIES
    VARIABLES=e12
    /ORDER ANALYSIS.

MEANS
    TABLES=e10 BY e12
    /CELLS MEAN COUNT STDDEV
    /STATISTICS ANOVA LINEARITY.

RECODE E12 (1=1)(2=2)(3=2)(6=3) INTO E12RECO.
VARIABLE LABELS E12RECO "Estado civil (recodificado en 3 categorías)".
VALUE LABELS E12RECO 1"Solteros" 2"Casados y conviven" 3"Viudos".
FREQUENCIES
    VARIABLES=e12 e12reco
    /ORDER ANALYSIS.
```

```
CROSSTABS
    /TABLES=a54reco BY e12reco
    /STATISTIC=CHISQ PHI
    /FORMAT= AVALUE TABLES
    /CELLS= COUNT COLUMN ASRESID
    /COUNT ROUND CELL.

FREQUENCIES
    VARIABLES=e10
    /ORDER ANALYSIS.

Recode e10 (18 thru 29=1) (30 thru 44=2) (45 thru 64=3) (65 thru 99=4) into EDAD.
Variable label EDAD "Edad (categorizada en cuatro categorías)".
Value labels EDAD 1"De 18 a 29 años (jóvenes)"
                  2"De 30 a 44 años (maduros)"
                  3"De 45 a 64 años (talludos)"
                  4"65 y más (mayores)".
FREQUENCIES
    VARIABLES=e10 EDAD
    /ORDER ANALYSIS.

CROSSTABS
    /TABLES=a54reco BY e12reco BY edad
    /STATISTIC=CHISQ PHI
    /FORMAT= AVALUE TABLES
    /CELLS= COUNT COLUMN ASRESID
    /COUNT ROUND CELL.

RECODE E12reco (3=sysmis).

CROSSTABS
    /TABLES=a54reco BY e12reco BY edad
    /STATISTIC=CHISQ PHI
    /FORMAT= AVALUE TABLES
    /CELLS= COUNT COLUMN ASRESID
    /COUNT ROUND CELL.
```

```
DO IF (edad > 2).
RECODE A54Reco (1=1)(2 3=2) INTO A54REC_2.
END IF.
EXECUTE.
VARIABLE LABELS A54REC_2 "GRADO DE FELICIDAD ACTUAL
    (RECODIFICADO EN 2 CATEGORÍAS)".
VALUE LABELS A54REC_2  1'POCO Y NADA FELIZ'
                       2'BASTANTE Y MUY FELIZ'.

CROSSTABS
    /TABLES=a54reco BY e12reco BY edad
    /STATISTIC=CHISQ PHI
    /FORMAT= AVALUE TABLES
    /CELLS= COUNT COLUMN ASRESID
    /COUNT ROUND CELL.

MEANS
    TABLES=e10 BY e12
    /CELLS MEAN COUNT STDDEV.

CROSSTABS
    /TABLES=edad BY E12reco
    /STATISTIC=CHISQ PHI
    /FORMAT= AVALUE TABLES
    /CELLS= COUNT COLUMN ASRESID
    /COUNT ROUND CELL.
```

Glosario

Análisis bivariable. Técnicas de análisis de datos que estudian simultáneamente dos variables.

Análisis multivariable o **multivariante.** Técnicas de análisis de datos que analizan simultáneamente más de dos variables.

Análisis univariable. Técnicas de análisis de datos que se ocupan de la distribución y los valores de una variable.

ANSI. Juego de caracteres básico, empleado por el entorno de trabajo Windows.

ASCII. (*American Standard Code for Information Interchange*). Código empleado para representar todos los caracteres alfanuméricos, signos de puntuación y control.

Asociación. "Término general usado para describir la relación entre dos variables. Esencialmente es sinónimo de correlación" (Everitt y Wykes, 2001: 28), si bien *asociación* se emplea más en el análisis de variables nominales y ordinales en tablas de contingencia.

Barra de herramientas. Conjunto de iconos, situados normalmente en la parte superior de cada ventana, que permiten acceder rápidamente a los procedimientos más usuales de trabajo, sin necesidad de acudir al menú y a los diferentes submenús.

CAPI (*Computer Assisted Personal Interview)*. Entrevistas personales asistidas por ordenador

CASI (*Computer Assisted Self Interviewing)*. Encuestas autorellenadas asistidas por ordenador

Categorizar variable. Proceso mediante el cual se transforma una variable cuantitativa en cualitativa. El SPSS realiza este proceso utilizando los percentiles, lo que permite que cada grupo contenga un número de casos similar.

CATI (*Computer Assisted Telephone Interview)*. Entrevistas telefónicas asistidas por ordenador

CAWI (*Computer Assisted Web Interview)*. Encuestas autorellenadas mediante web.

Celdilla. Unión de una variable fila y una variable columna.

Codificación. Clasificar sistemática de las respuestas en categorías exhaustivas y mutuamente excluyentes, con el fin de trasladar la información a un suporte informático.

Correlación parcial. Relación entre dos variables eliminando la influencia de una tercera

Correlación. Término general usado para describir la relación entre dos variables. Normalmente se emplea para describir la relación entre dos variables cuantitativas, métricas.

Cuartiles. Valores que dividen una distribución de frecuencias en cuatro partes iguales.

Depuración (de datos). Conjunto de técnicas que permiten, a partir de la información recogida en la encuesta o utilizando otra información adicional, corregir la mayor parte de los errores de la encuesta.

Depuración por contraste. Sistema de validación que consiste, básicamente, en realizar una grabación por duplicado (y en ocasiones empelando personas distintas) para después considerar las diferencias existentes entre ambos archivos.

Desviación típica. Medida de dispersión de un conjunto de observaciones. Es la raíz cuadrada del sumatorio de las diferencia entre cada valor y la media de la serie.

Distribución de frecuencias. División de una muestra de observaciones en un conjunto de clases, junto con el número de observaciones de cada clase. Actúa como un útil resumen de las principales características de los datos, como son la localización, la forma y la dispersión (Everitt y Wykes, 2001: 69).

Distribución de frecuencias. División de una muestra de observaciones en un conjunto de clases, junto con el número de observaciones de cada clase. Actúa como un útil resumen de las principales características de los datos, como son la localización, la forma y la dispersión (B.S. Everitt y T. Wykes, *Diccionario de estadística para psicólogos*. Barcelona: Ariel, 2001, pág. 69).

Distribución de frecuencias. División de una muestra de observaciones en un conjunto de clases, junto con el número de observaciones de cada clase. Actúa como un útil resumen de las principales características de los datos, como son la localización, la forma y la dispersión (B.S. Everitt y T. Wykes, *Diccionario de estadística para psicólogos*. Barcelona: Ariel, 2001, pág. 69).

Edición. Proceso en el que las respuestas son inspeccionadas, corregidas y en ocasiones precodificadas de acuerdo a un conjunto de reglas fijas

Editor de datos SPSS. Menú principal de SPSS, donde se presentan los menús del programa y la ventana de datos activos.

Filtrar. Seleccionar los casos que cumplen unas determinadas características, para realizar los análisis posteriores excluidos estos casos.

Imputación. Dar nuevos valores a los datos faltantes en una variable, valores que cumplan determinadas condiciones.

Inconsistencias (entre preguntas). Respuestas diferentes de un mismo entrevistado en preguntas similares del cuestionario.

Índice de relación Parcial. Índice que muestra la relación entre dos variables (X e Y), eliminando la influencia de una tercera (Z).

Índice. Variable formada por la unión de varias variables.

Informe. Documento donde se presentan todos los resultados de la investigación.

Libro de códigos. Documento donde se recogen todas las posibilidades de respuesta de una pregunta (o registro), identificados con el número correspondiente. En el apartado 3.14 se muestra el libro de códigos del cuestionario presentado al final de este texto.

Marco muestral (marco de muestreo). Documento donde queda recogido el universo de la investigación, toda la población objeto de estudio.

Marginal (distribución marginal). Distribución de casos de una variable, frecuencia de una variable.

Media aritmética. Promedio de una distribución. Es la suma de valores de una distribución de frecuencias, dividida entre el número de casos.

Mediana. Valor central de la distribución, valor que divide la distribución de frecuencias en dos partes de igual tamaño.

Medida de asociación. Índice numérico que indica la existencia, grado y dirección de la relación entre dos variables

Moda. Valor más frecuente en una distribución de frecuencias.

Neutralización. Proceso mediante el cual se *controla* el efecto de una variable, aumentando así el poder de las *variables explicativas* (ver variable explicativa, variable interviniente, variable controlada, variable no controlada, y variable perturbadora).

Nivel de significación. "Nivel de probabilidad en el que se acuerda que se rechazará la hipótesis nula. De forma convencional se fija en 0,05" (Everitt y Wykes, 2001: 150).

No respuesta. Ausencia de respuesta por parte de una unidad.

No respuesta parcial. Ausencia de respuesta producida porque el entrevistado que estaba contestando el cuestionario ha decidido no responder a determinadas cuestiones.

No respuesta total. Ausencia de respuesta producida porque una persona no ha contestado ninguna pregunta del cuestionario.

Objetivo general. Enunciado claro y preciso de las metas que se persiguen con la investigación.

Objetivos específicos. Indican lo que se pretende lograr en cada una de las etapas de la investigación, implicando así un mayor nivel de concreción temporal, temática y estratégica

OCR (*Optical Character Recognition)*. Reconocimiento óptico de caracteres.

OMR (*Optical Mark Recognition)*. Reconocimiento mecánico de caracteres.

Paquete estadístico. Programa informático que realiza de forma integrada operaciones variadas como modificar datos, crear índices y escalas, analizar datos, etc.

Ponderar. Proporcionar a los casos diferente "pesos" para el análisis estadístico, elaborando una *replica simulada* de determinados casos que aumenta (o disminuye) su valor en el conjunto de la muestra.

Pregunta contingente. Preguntas *filtradas*, preguntas cuya respuesta depende de la respuesta de una pregunta filtro.

Pregunta filtro: Preguntas utilizadas para seleccionar sujetos que cumplan unas determinadas características, y dirigir a este subcolectivo un conjunto de preguntas del cuestionario. Una vez que los entrevistados han demostrado un determinado nivel de conocimientos se les pide una opinión sobre el tema, evitando que los consultados expresen opiniones sobre asuntos que no conocen.

Preguntas de respuesta múltiple, preguntas multirespuesta. Preguntas con categorías de respuesta no excluyentes, y que solicitan del entrevistado más de una respuesta.

Rango. Medida de dispersión de un conjunto de observaciones. Se calcula restando al valor más alto de la distribución el valor más bajo.

Recodificación. Agrupación, unión de categorías.

Relación (asociación) condicional. Asociación entre variables de tablas parciales, relación entre dos variables eliminando la influencia de una tercera.

Relación (asociación) marginal. Relación entre dos variables sin tener en cuenta ninguna otra.

Relación no directa, relación intervenida. Existencia de una *aparente* relación entre dos variables, que se realidad tiene lugar por la influencia de una tercera.

Residuo/residual tipificado. Valor residual (residuo), eliminando el efecto del número de casos.

Residuo/residual. Diferencia entre la frecuencia observada y la esperada (o teórica).

SPSS. (*Statistical Package for Social Sciences)*. Paquete estadístico para las Ciencias Sociales.

Supervisión (del trabajo de campo). Proceso de comprobación de que la información se ha recogido –formal y materialmente– de acuerdo a las instrucciones previamente distribuidas por los directores del trabajo y por los responsables de la red de entrevistadores (Wert, 1996: 56).

Tabla cuatridimensional. Tabla de cuatro dimensiones

Tabla tridimensional. Tabla de tres dimensiones

Trabajo de campo. "Aplicación del cuestionario a la muestra, es decir, la recogida de la información mediante aplicación del cuestionario con algunos de los diferentes procedimientos existentes" (Alvira, 2004: 15).

Universo, población. Conjunto total de unidades que van a ser estudiadas en una investigación. Ante la imposibilidad de estudiarlas todas normalmente se selecciona una parte que es definida como *muestra*.

Everitt y Wykes proporcionan la siguiente definición de *población*: "término utilizado para cualquier colección finita o infinita de 'unidades', que a menudo son personas, pero pueden ser, por ejemplo, instituciones, sucesos, etc". (B.S. Everitt y T. Wykes, *Diccionario de estadística para psicólogos*. Barcelona: Ariel, 2001, pág. 155).

Valores atípicos. Observaciones incluidas en la matriz de datos que muestran inconsistencias con el resto de la distribución

Variable interviniente, variable extraña. Variables ajenas a la relación causa-efecto buscada con la investigación (Díaz de Rada, 2002: 24).

Variables *básicas* de la muestra. Reciben este nombre las variables que han sido utilizadas para la selección muestral, las variables que son consideradas para verificar que la muestra se *asemeja* a la población (estas variables presentan la misma distribución en la muestra que en la población).

Variables controladas. Variables extrañas que pueden ser controladas por el investigador. Este control puede realizarse *a priori* por medio del diseño de investigación, o *a posteriori* mediante al utilización de determinadas técnicas de análisis de datos (Díaz de Rada, 2002: 24).

Variables de identificación. Variables que identifican a los entrevistados; también conocidas por variables sociodemográficas, variables independientes. En algunos contextos se habla de *cabeceras*, en la medida que son las variables fijas que aparecen en todas las tablas.

Variables explicativas. Variables que constituyen el objetivo de la investigación, variables que pretendemos medir o recoger. Dentro de éstas es posible diferenciar entre variables independientes o *predictoras* (X) que recogen la causa de la explicación, y variables dependientes o *pronosticadas*, definidas como el efecto (Y) producido por las variables anteriores (Díaz de Rada, 2002: 24).

Variables no controladas. Variables extrañas que no son controladas por el investigador. Forman parte de este grupo las variables aleatorizadas y las perturbadoras (Díaz de Rada, 2002: 24).

Variables perturbadoras. Variables extrañas no controladas que pueden afectar a las variables explicativas (Díaz de Rada, 2002: 24).

Ventana de datos. Parte central del editor de datos de SPSS, donde se presentan en filas la información de cada caso, y en las columnas cada una de las variables.

Bibliografía

Abascal, E.; Grande, I. (2005). *Análisis de Encuestas*. Madrid: ESIC.

Alaminos, A. (1993). *Gráficos*. Madrid: Centro de Investigaciones Sociológicas, colección Cuadernos Metodológicos, número 7.

Alvira, F.; Blanco, F. (2000). "Introducción al análisis de datos", en M. García Ferrando; J. Ibáñez y F. Alvira (Eds.), *El Análisis de la Realidad Social*. Madrid: Alianza Universidad, pp. 485-524.

Alvira, F. (2004). *La encuesta: una perspectiva general metodológica*. Madrid: CIS, Cuadernos Metodológicos, nº 35.

ANEIMO (2000). "Estándar de Calidad en la Investigación de mercados (ECIM)", en ESOMAR, *Códigos y guías de ESOMAR y normas aplicables a la investigación de mercados*. Barcelona: AEDEMO.

Aparicio, F. (1991). *Tratamiento informático de encuestas*. Madrid: Ra-ma.

Ayerdi, P. (1995). *Estilos de Vida y Desigualdad Social en España*, Tesis Doctoral. Pamplona: Universidad Pública de Navarra

Azofra, M. J. (1999). *Cuestionarios*. Madrid: CIS, Cuadernos Metodológicos, nº 26.

Calvo, F. (1990). *Estadística Aplicada*. Bilbao: Deusto.

Camarero Rioja, A. (2002). "Acerca de las medidas de asociación en investigación social: un viejo problema que conviene no olvidar", en J. M. Arribas Macho y M. Barbur (coordinadores), *Estadística y Sociedad*. Madrid: UNED, pp. 377-397.

Cea D'Ancona, M. A. (2002). Análisis Multivariable. Teoría y práctica en la investigación social. Madrid: Síntesis.

Cea D'Ancona, M. A. (2004). *Métodos de encuesta. Teoría y práctica, errores y mejora*. Madrid: Síntesis.

Díaz de Rada, V. (1999). *Técnicas de Análisis de Datos para Investigadores Sociales*. Madrid: RaMa.

Díaz de Rada, V. (2000). *Problemas originados por la no respuesta en investigación social: Definición, control y tratamiento*. Pamplona: Universidad Pública de Navarra.

Díaz de Rada, V. (2001). *Diseño y elaboración de cuestionarios para la investigación comercial*. Madrid: Esic.

Díaz de Rada, V. (2002). *Técnicas de Análisis Multivariante en investigación social y comercial*. Madrid: RaMa.

Díaz de Rada, V. (2004). *Estrategias de consumo y estilos de vida en la sociedad navarra del siglo XXI*. Pamplona: Universidad Pública de Navarra.

Dometrius, N. C. (1992). *Social Statistics using SPSS*. Londres: Harper-Collins Pulishers.

ESOMAR (2000). *Códigos y guías de ESOMAR y normas aplicables a la investigación de mercados*. Barcelona: AEDEMO.

Everitt, B. S.; Wykes, T. (2001). *Diccionario de estadística para psicólogos*. Barcelona: Ariel.

Felman, A. *et al.* (1989). "La Estructura Social y el Apoyo Partidista en España", *Revista Española de Investigaciones Sociológicas*, vol. 47, pp. 7-72

García Ferrando, M. (1985). *Socioestadística*. Madrid: Alianza Universidad.

González, J. J. (1994). "Sobre el Declive Político de las Clases", *Economía y Sociedad*, Vol. 11, diciembre, pp. 9-24.

Grande Esteban, I.; Abascal Fernández, E. (1999). *Fundamentos y técnicas de investigación comercial*. Madrid: ESIC (Edición original 1994).

Granero, R.; Doménech, J. M.; Bonillo, A. (2001). "Estudio de la eficacia de la técnica de verificación aleatoria como alternativa a la doble entrada de los datos: ventajas y limitaciones", en *Metodología de Encuestas*, vol. 3, nº 1, pp. 1-13.

Haberman, S. J. (1973). "The Analysis of Residuals in Cross-Classified Tables", *Biometrics*, vol. 29, pp. 205-220.

López Pintor, R.; Wert, J. I. (2000). "El análisis de los datos de encuesta", en M. García Ferrando, J. Ibáñez y F. Alvira (eds.), *El Análisis de la Realidad Social*. Madrid: Alianza, 3ª ed., pp. 525-554. (Edición original 1986).

Luque Martínez, T. (Coordinador) (2000). *Técnicas de Análisis de Datos en Investigación de Mercados*. Madrid: Pirámide.

Manzano, V. (1995). *Inferencia Estadística: Aplicaciones con SPSS/PC+*. Madrid: Rama.

Mejías, G. (2005). "Las incidencias", en Instituto Nacional de estadística, *Los Trabajos de Campo en las encuestas del INE*. Curso impartido en la Escuela de Estadística de las Administraciones públicas, del Instituto Nacional de Estadística. 21–23 de junio.

Miguel, A. de (1997). *Manual del perfecto sociólogo*. Madrid: Espasa Calpe.

Padilla García, J. L.; González, A.; Pérez, C. (1998). "Elaboración del cuestionario", en A. J. Rojas, J. S. Fernández y C. Pérez (Eds.), *Investigar mediante encuestas*. Madrid: Síntesis, pp. 115-140.

Reynolds, H. T. (1984). *Analysis of Nominal Data*. California: Sage University Papers Series on QASS.

Rojas Tejada, A. J.; Fernández Prados, J. S. (2000). "Introducción al tratamiento de datos", en A. J. Rojas Tejada; J. S. Fernández Prados y C. Pérez Meléndez (Eds.), *Investigar Mediante encuestas*. Madrid: Síntesis, pp. 169-177.

Sánchez Carrión, J. J. (1999). *Manual de análisis estadístico de los datos*. Madrid: Alianza.

Sánchez Carrión, J. J. (2000). *La calidad de la encuesta: el caso de la no respuesta*. Madrid: Alianza.

Ruiz-Maya, L. *et al.* (1990). *Metodología estadística para el Análisis de Datos Cualitativos*. Madrid: ICO-CIR.

Sevilla Moróder, J. (2006). *Gramática de las gráficas. Pistas para mejorar las representaciones de datos*. Pamplona: Universidad Pública de Navarra.

Villan Craido, I.; Bravo Cabria, M. S. (1990). *Procedimiento de Depuración de Datos Estadísticos*, Vitoria: Euskadi.

Villarejo Ramos, A. F. *et al.* (1996). "Estrategias de conversión entre escalas de medición mixtas aplicables en la investigación de marketing", *VIII Encuentros de Profesores Universitarios de Marketing*, pp. 465-478.

Rodriguez Osuna, J.; Ferreras, M. L.; Núñez, A. (1991). "Inferencia Estadística, Niveles de Precisión y Diseño Muestral", *Revista Española de Investigaciones Sociológicas*, vol. 54, pp. 139-162.

Sánchez Carrión, J. J. (2000). *La calidad de la encuesta: el caso de la no respuesta*, Madrid: Alianza.

Stevens, S. S. (1946). "On the Theory of Scales of Measurement", *Science*, vol. 103, pp. 677-680.

Wert, J. I. (1996). *Carta abierta a un incrédulo sobre las encuestas y su muy disputado crédito*. Madrid: Península.